Multisensory Interaction in Virtual Reality with Embodied Conversational Agent

Dissertation submitted to
Visvesvaraya National Institute of Technology, Nagpur
in partial fulfillment of the requirements for the award of
the degree

Master of Technology
in
Computer Science & Engineering

by

GOUTAM GOPE (MT23MCS020)

under the guidance of

***Dr. Shital A. Raut,**

”Dr. Mukesh Barange, ”Dr. Beatrice Biancardi

***Department of Computer Science & Engineering**

Visvesvaraya National Institute of Technology

Nagpur 440 010 (India)

In collaboration with

”LINEACT CESI, Engineering School,

Nanterre 92000 (France)

JUNE 2025

Multisensory Interaction in Virtual Reality with Embodied Conversational Agent

Dissertation submitted to
Visvesvaraya National Institute of Technology, Nagpur
in partial fulfillment of the requirements for the award of
the degree

Master of Technology
in
Computer Science & Engineering

by

GOUTAM GOPE (MT23MCS020)

under the guidance of

***Dr. Shital A. Raut,**

”Dr. Mukesh Barange, ”Dr. Beatrice Biancardi

***Department of Computer Science & Engineering**

Visvesvaraya National Institute of Technology

Nagpur 440 010 (India)

In collaboration with

”LINEACT CESI, Engineering School,

Nanterre 92000 (France)

JUNE 2025

Department of Computer Science & Engineering
Visvesvaraya National Institute of Technology, Nagpur

Declaration

I, **Goutam Gope**, hereby declare that this dissertation titled **" Multisensory Interaction in Virtual Reality with Embodied Conversational Agent "** is carried out by me in the Department of Computer Science & Engineering of Visvesvaraya National Institute of Technology, Nagpur. The work is original and has not been submitted earlier whole or in part for the award of any degree at this or any other Institution.

Goutam Gope

Name and Signature

Date:

Certificate

This to certify that the dissertation titled **" Multisensory Interaction in Virtual Reality with Embodied Conversational Agent "** is submitted by **Goutam Gope** in partial fulfillment of the requirements for the award of the degree of **Master of Technology in Computer Science & Engineering**, VNIT Nagpur. The work is comprehensive, complete, and fit for final evaluation.

12/6/25

Dr. Shital A. Raut
Department of Computer Science & Engineering, VNIT Nagpur

सह. प्राध्यापक
Associate Professor
संगणक विज्ञान एवं अभियांत्रिकी विभाग
Dept of Computer Science & Engg
वि रा प्रौ स नागपुर-10 (भारत)
NAGPUR 0010 (India)

Dr. M. P. Kurhekar
Head, Department of Computer Science & Engineering, VNIT Nagpur
Date:

विभाग प्रमुख
Head
संगणक विज्ञान एवं अभियांत्रिकी विभाग
Dept. of Computer Science & Engg.
वि.रा.प्रौ.सं. नागपुर-१० (भारत)
V.N.I.T., Nagpur-440010. (India)

Department of Computer Science & Engineering
Visvesvaraya National Institute of Technology, Nagpur

Non-Plagiarism Certificate

Certified that M. Tech Dissertation titled **"Multisensory Interaction in Virtual Reality with Embodied Conversational Agent "** submitted by **Goutam Gope**, Enrollment No. **MT23MCS020**, ID No. **30713** has been checked for plagiarism using TURNITIN software and the overlap is found within prescribed limits. A summary of the report is given below.

Thesis_Goutam_Gope.pdf

ORIGINALITY REPORT

9%	7%	5%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

Goutam Gope

Goutam Gope
Enrollment No: MT23MCS020

12/6/25

Dr. Shital A. Raut
Associate Professor
Department of Computer Science & Engineering
VNIT, Nagpur

सह. प्राध्यापक
Associate Prof.,SSOI
संगणक विज्ञान एवं अभियांत्रिकी विभाग
Dept of Computer Science & Engg
वि.रा.प्रौ.सं. नागपुर-10 (भारत)
VNIT NAGPUR 0010 (India)

ACKNOWLEDGEMENT

I would like to express my deepest gratitude to all who have supported and guided me throughout my M.Tech journey.

My sincere thanks to my supervisor, **Dr. Shital A. Raut**, for her exemplary guidance, steadfast support, and insightful feedback, which have been fundamental to my academic growth. I also thank **Dr. Mukesh Barange** and **Dr. Beatrice Biancardi** for their valuable insights and encouragement that greatly enhanced this work.

I am grateful to **Dr. M. P. Kurhekar**, Head of Department, for his invaluable support and motivation throughout this project.

The faculty members of the Department of Computer Science & Engineering at Visvesvaraya National Institute of Technology Nagpur are also to be thanked for their assistance, guidance, and knowledge in securing the resources required for this research.

I would also like to express my gratitude to my Family and Friends for their unwavering support, understanding, and encouragement throughout this journey. Their love, motivation, and belief in my abilities have been the driving force behind my accomplishments.

I also offer my silent reverence to the One whose fierce grace watches over me, guiding me through shadows into light. Her presence needs no name—only devotion.

Thank you all.

Goutam Gope

ABSTRACT

The intricacy of comprehending human emotions in virtual settings is frequently hindered by the dependence on unimodal inputs such as facial expressions. With the use of a machine learning model that combines three different inputs—text, video, and audio—this thesis offers a unique method for emotion identification that improves prediction accuracy by taking into account contextual subtleties. The approach overcomes the drawbacks of unimodal systems by integrating textual content, facial clues, and vocal tone, allowing for a more comprehensive understanding of emotional states. The study also investigates how this model might be used in virtual reality (VR) environments, particularly with embodied conversational agents—a type of agent that is made to engage with humans in a natural way.The thesis provides an introduction analysis of VR settings and embodied agents to contextualize their significance, even though the major focus is still on creating and assessing the multi-input emotion detection model. The results provide a basis for future integration of multimodal emotion detection in immersive, interactive systems, however a thorough investigation of VR implementation and agent design is outside the purview of this work. By highlighting the importance of context-aware, multimodal approaches in comprehending emotions, this research advances human-computer interaction.

TABLE OF CONTENTS

Title **Page No.**

ABSTRACT . . . . . **iv**

TABLE OF CONTENTS . . . . . **v**

LIST OF TABLES . . . . . **ix**

LIST OF FIGURES . . . . . **x**

1 Introduction **1**

1.1 Background and Motivation 1

1.2 Problem Statement 2

1.3 Objectives of the Thesis 3

2 Literature Survey **4**

2.1 Foundations of Emotion Recognition 5

2.1.1 Theories of Human Emotion in Psychology and Cognitive Science 5

2.1.2 Importance of Emotions in Human Communication 7

2.1.3 Tactile Emotion Communication in Human Psychology 7

2.2 Traditional Approaches to Emotion Detection 8

2.2.1 Unimodal Emotion Recognition Techniques 8

2.2.2 Limitations of Unimodal Approaches 10

2.3 Emergence of Multimodal Emotion Recognition 10

2.3.1 Motivation for Multimodal Approaches 11

2.3.2 Fusion Techniques 11

2.3.3 Benchmark Datasets 12
2.4 Deep Learning in Emotion Detection 13
2.4.1 Benefits of Deep Learning in Emotion Recognition 14
2.4.2 Cross-Modality Applications 14
2.4.3 Challenges in Deep Learning-Based Emotion Recognition 15
2.5 Emotion Detection in Virtual Reality (VR) 16
2.6 Embodied Conversational Agents (ECAs) 16
2.7 Multimodal Emotion Detection for ECAs in VR 17
2.7.1 Haptic Feedback and Remote Touch in Emotion Communication . 17
2.8 Context-Aware Emotion Detection 18

3 Implementation 19
3.1 Datasets 19
3.1.1 Data Preprocessing 19
3.1.2 Dataset Structure and Assumptions 20
3.2 Text Preprocessing 21
3.2.1 Tokenization with BERT 21
3.2.2 Truncation and Padding 21
3.2.3 Encoding to Numerical IDs 21
3.2.4 Significance 22
3.3 Audio Preprocessing 22
3.3.1 Loading and Validating Audio 22
3.3.2 Standardizing Audio Length 22
3.3.3 Output Format 23
3.3.4 Significance 23
3.4 Video Preprocessing 23
3.4.1 Video Modality 23
3.4.2 Loading Video Frames 23
3.4.3 Applying Transformations 24
3.4.4 Output Format 24
3.4.5 Significance 25

3.5 Extracted Face Preprocessing 25
3.5.1 Loading and Validating Face Data 25
3.5.2 Loading Main Face Data 25
3.5.3 Loading Extracted Face Data 26
3.5.4 Significance 27
3.5.5 Dataset and Dataloading 29
3.6 Model Architecture 30
3.6.1 Overview of the Multimodal Architecture 31
3.6.2 Shared Modality Encoders 32
3.6.3 Video Encoder 34
3.6.4 Face Encoder 36
3.6.5 Multimodal Fusion Mechanism 37
3.6.6 Classification Heads 39
3.6.7 Training Considerations 40
3.6.8 Significance of the Architecture 41

4 Experiments and Results 43
4.1 Text-Only Modality 43
4.1.1 Test Set Evaluation Results 43
4.2 Text + Audio Modalities 44
4.2.1 Fusion Model Test Results 44
4.3 Text + Video Modalities 45
4.3.1 Fusion Model Test Results 45
4.4 Audio + Video Modalities 45
4.4.1 Fusion Model Test Results 45
4.5 Text + Audio + Video Modalities 46
4.5.1 Fusion Model Test Results 46
4.6 Text + Audio + Video + face Modalities with ViT/VideoMAE 47
4.6.1 Fusion Model Test Results 47
4.7 Text + Audio + Video + face Modalities with Convanxt 49
4.7.1 Fusion Model Test Results 49

4.8 Comparative Analysis of ConvNeXtEnhancedFRAM and ViT/VideoMAE Architectures . . . 51
4.8.1 Overview . . . 51
4.8.2 Architectural Benefits . . . 51
4.8.3 Correlation with MELD Dataset . . . 52
4.8.4 Advantages of Training and Optimization . . . 52
4.8.5 Empirical Efficacy . . . 53
4.8.6 Conclusion . . . 53
4.9 Comparison with State of the Art Model . . . 54

5 Discussion . . . 55
5.1 Discussion . . . 55

6 Conclusion And Future Work . . . 57
6.1 Conclusion . . . 57
6.2 Future Work . . . 58

List of Tables

4.1 Text Only Classification Report 44
4.2 Text + Audio Classification Report 44
4.3 Text+Video Classification Report 45
4.4 Audio+Video Classification Report 46
4.5 Text+Audio+Video Classification Report 47
4.6 Text+Audio+Video+Face Classification Report with ViT/VideoMAE . . 48
4.7 Text+Audio+Video+face Modalities with ConvNeXt 50
4.8 Performance Comparison Between Models Using Text, Face, Video, and Voice Modalities 54

List of Figures

3.1 Architecture of Model A 32
3.2 Architecture of Model B 32
3.3 bert-base-uncased Text Encoder 33
3.4 wav2vec2-base-960h Audio Encoder 33
3.5 Video Encoder of Model A 35
3.6 Video Encoder of Model B 35
3.7 Face Encoder of Model A 36
3.8 Face Encoder of Model B 37
3.9 Multimodel Integration of Model A 38
3.10 Multimodal Integration of Model B 39
3.11 Classification Mechanism of Model A 40
3.12 Classification Mechanism of Model B 40

4.1 Text + Audio + Video Modalities Confusion Matrix 48
4.2 Text+Audio+Video+Face with Action unit model validation Accuracy Graph 49
4.3 Enter Caption 50
4.4 Text+Audio+Video+Face with convanxt validation Accuracy Graph . . . 51

Chapter 1

Introduction

1.1 Background and Motivation

Psychology, cognitive science, and more recently, human-computer interaction (HCI), have long placed a high priority on comprehending human emotions. Machines that can reliably and instantly understand human emotions are becoming more and more necessary as technology advances, especially with the emergence of immersive environments like virtual reality (VR). People's communication, decision-making, and social interactions are greatly influenced by their emotions. As a result, creating systems that are sensitive to human emotions could greatly improve user experience across a range of applications, including virtual assistants, immersive learning environments, and mental health treatments.

In the past, emotion identification algorithms have mostly relied on unimodal inputs, such as speech tones or facial expressions. Even while these approaches have had some success, they frequently fail in intricate, real-world situations where emotional cues could be contradicting, nuanced, or context-dependent. A user may, for example, vocally indicate enthusiasm while displaying signals of discomfort on their face, or their vocal intonation may convey ambiguity that cannot be resolved by a single modality. These drawbacks emphasize the necessity of a more comprehensive, situation-specific strategy.

New possibilities for integrating various input modalities to gain a deeper understanding of human affect have been made possible by recent developments in machine learning and deep learning. Textual content, video facial expressions, and audio vocal

characteristics can all be combined to create systems that are more reliable, accurate, and sensitive to the subtleties of emotional expression. In addition to improving prediction accuracy, multimodal emotion detection indicates a more human-like approach to perception, taking into account several senses and information sources at once.

In the context of embodied conversational agents (ECAs), which are virtual agents created to mimic in-person interactions with people, this research is very pertinent. The ability of ECAs to identify and adjust to a user's emotional state becomes crucial in virtual reality (VR) situations, where users are submerged in rich, interactive landscapes. However, the multimodal and contextual richness needed for these situations is frequently absent from current emotion identification algorithms. For virtual interactions to become more organic, sympathetic, and successful, this gap must be filled.

In the context of embodied conversational agents (ECAs), which are virtual agents created to mimic in-person interactions with people, this research is very pertinent. The ability of ECAs to identify and adjust to a user's emotional state becomes crucial in virtual reality (VR) situations, where users are submerged in rich, interactive landscapes. However, the multimodal and contextual richness needed for these situations is frequently absent from current emotion identification algorithms. For virtual interactions to become more organic, sympathetic, and successful, this gap must be filled.

1.2 Problem Statement

Current unimodal emotion detection algorithms, focusing on facial expressions, speech, or text, fail to capture the complex, context-dependent character of human emotions in dynamic VR scenarios with embodied conversational agents (ECAs). The absence of comprehensive multimodal models combining textual, visual, aural, and facial inputs, while addressing data discrepancies and class imbalances, inhibits emotionally intelligent interactions. This thesis proposes to design a multimodal emotion detection model to boost accuracy and contextual comprehension for VR-based ECAs, overcoming unimodal limits for realistic, sympathetic virtual interactions.

1.3 Objectives of the Thesis

This thesis aims to improve the accuracy and contextual comprehension of emotional states by creating and assessing a multimodal emotion detection model that integrates textual, visual, and aural input. The study examines the applicability and potential use of such a system in VR environments, especially including ECAs, even if the primary focus is still on model creation and evaluation.

Despite not delving further into the technical issues of designing agent behavior or implementing VR systems, this thesis offers a conceptual framework for further study in these fields. The goal is to establish the foundation for incorporating emotion-aware algorithms into immersive settings that facilitate more natural and human-like interactions.

This work adds to the ongoing attempts to develop intelligent systems that can more accurately and empathetically interpret and react to human emotions by addressing the shortcomings of unimodal systems and highlighting the need of contextual understanding in emotion detection.

Chapter 2

Literature Survey

This chapter provides a thorough assessment of both fundamental and recent research pertinent to emotion identification, especially in the context of multimodal systems and their applications in virtual worlds. We start by examining the psychological and cognitive scientific foundations of human emotion, addressing current ideas and the importance of emotions in everyday communication. This establishes the foundation for comprehending the computational modeling of emotional states.

We then analyze conventional methods of emotion identification, concentrating on unimodal techniques that use a single source of information, such as facial expressions, voice, or text. The shortcomings of these methods underscore the need for multimodal emotion identification, which integrates several sensory inputs to enhance detection accuracy and robustness.

Subsequently, we explore the progression of deep learning in emotion detection, detailing its benefits for feature extraction and scalability, as well as the obstacles it presents. We investigate the integration of emotion detection into upcoming technologies such as Virtual Reality (VR) and Embodied Conversational Agents (ECAs), highlighting the need for real-time, context-sensitive emotional comprehension in interactive systems.

The chapter ultimately examines contemporary initiatives to integrate haptic feedback and tactile communication into emotion-aware systems, and analyzes how contextual clues might improve the accuracy and profundity of emotion identification. This literature study establishes the conceptual and technological foundation for the design and imple-

mentation of our proposed multimodal emotion detection system, particularly in virtual reality settings with embodied conversational agents.

2.1 Foundations of Emotion Recognition

Our knowledge of how emotions are created, expressed, and perceived is the basis of emotion recognition. Emotions are fundamental to the human experience, impacting everything from memory and decision-making to social interaction and education. Theoretical models that assist explain the complexity of human affect have been developed during decades of research into the mechanisms underlying emotional reactions by fields such as psychology and cognitive science.

In the field of artificial intelligence, these fundamental discoveries are increasingly proving crucial, particularly when developing systems that seek to communicate with people in a natural and intuitive manner. As technology becomes more and more integrated into daily life—through social robots, virtual assistants, and immersive settings like virtual reality—machines must be able to identify and react to human emotions in addition to understanding commands.

Conventional methods of emotion identification concentrated on individual indicators, including tone of voice or facial expressions. But the subtleties of emotional conduct in the real world frequently call for a more comprehensive and complete knowledge. Researchers are attempting to create more intelligent, sympathetic, and context-aware computers by acknowledging the complexity of emotional expression and establishing computational models in psychological theories.

The theoretical foundation of emotions, their function in communication, and how these ideas form the basis of contemporary emotion recognition systems are all covered in this part.

2.1.1 Theories of Human Emotion in Psychology and Cognitive Science

Two of the most important frameworks for tackling the difficult problem of comprehending human emotions are Robert Plutchik's psychoevolutionary model of emotions and Paul Ekman's theory of fundamental emotions. Ekman proposed six emotions—happiness,

sorrow, fear, anger, surprise, and disgust—that are universally recognizable across cultures through particular facial expressions in his 1992 work, "An Argument for Basic Emotions." These emotions are physiologically hardwired rather than culturally learnt, as his cross-cultural research—which included investigations among isolated and illiterate societies—showed (Ekman, 1992). Different facial muscles move spontaneously and automatically in response to each basic emotion. According to Ekman, these feelings are usually fleeting, strong, and have adaptive purposes linked to survival—disgust, for instance, aids in avoiding dangerous substances, while fear prompts the avoidance of hazards. Furthermore, even in the absence of spoken language, facial expressions serve as social signals that convey inner emotional states.Ekman's model is fundamental to comprehending the expression and perception of emotions due to its universality and biological underpinning (Ekman, 1992).

Adding to the complexity of emotions, Robert Plutchik's 1980 psychoevolutionary theory offered a multifaceted framework depicted as a wheel of emotions. To show how emotions mix and fluctuate in intensity, he devised a circular model that incorporated eight main feelings: joy, trust, fear, surprise, sadness, disgust, wrath, and anticipation (Plutchik, 1980). According to Plutchik's paradigm, emotions are like colors, and basic emotions can combine to create more complicated sentiments. For example, love might arise from joy and trust, whereas aggression can arise from anticipation and fury. According to his idea, emotions can range in intensity from moderate to intense (e.g., fury to annoyance) and exist along continuous dimensions rather than just discrete categories.The vast range of human emotional experiences and their adaptive functions in directing action, such as approach, avoidance, or defense, are explained by this dynamic and multi-layered approach (Plutchik, 1980).

These models work well together because they offer a biological basis as well as a sophisticated framework for comprehending the complexity and fluidity of human emotions. The distinct, quantifiable emotional categories provided by Ekman's work have been helpful in affective computing and human-computer interaction. Plutchik's wheel, on the other hand, enhances this comprehension by illustrating how emotions fluctuate and merge in actual circumstances. These observations serve as the foundation for con-

temporary multimodal emotion recognition systems, which seek to more sensitively and accurately decipher emotions from voice, facial expressions, and contextual signals.

2.1.2 Importance of Emotions in Human Communication

Human connection, communication, and understanding of the world are all heavily influenced by emotions. Subtle clues like tone of voice and facial expressions frequently influence how a message is interpreted long before words are comprehended. Albert Mehrabian (1971) highlighted the importance of nonverbal communication in expressing emotional meaning in his seminal study. People rely much more on how something is expressed (tone: 38%) and what is seen (facial expressions: 55%) than on the actual words said (only 7

At the same time, Antonio Damasio's 1994 research expanded our comprehension of emotions by adding a neuroscientific perspective. Descartes demonstrated in Error how closely human logic and decision-making are entwined with emotions. Even when logical reasoning is unaffected, his studies on people with brain injuries showed that emotional areas in the brain can be disrupted, making even basic decisions challenging. This offered compelling evidence that emotions do not obstruct rational reasoning as was previously supposed, but rather aid in decision-making, memory, and social interaction.

Taking these viewpoints together, it is evident that emotions are more than just passing sensations; they are necessary for meaningful interactions, successful communication, and well-informed decision-making. Designing emotionally intelligent systems requires the ability to identify and analyze emotional cues from a variety of modalities, such as tone, expression, and context. This is especially important in immersive or socially engaging settings like virtual reality.

2.1.3 Tactile Emotion Communication in Human Psychology

Although affective computing and psychology theories have traditionally focused on facial and verbal expressions of emotion, Hertenstein et al. (2006) investigated touch, a less common but incredibly human medium. They conducted three controlled tests to determine whether physical touch alone is an accurate means of identifying and commu-

nicating emotional states.

Without verbal or visual cues, individuals in their studies who were touched on the arm were able to correctly identify emotions including anger, fear, disgust, love, gratitude, and pity at far higher-than-chance levels. Furthermore, different touch behaviors (such as patting, squeezing, and stroking) were consistently associated with particular emotions, indicating that touch is a modality that can communicate diverse emotional states rather than just being a generalized emotional amplifier.

Their conclusions also held true for viewers of tactile interaction videos. The expressive richness and social universality of touch as a channel of emotional interaction were highlighted by the viewers' ability to accurately decipher the intended mood even when they were not touched themselves.

Modern affective computing systems are encouraged to think about going beyond the conventional audio-visual paradigm by this study, which supports the notion that emotion communication is intrinsically multimodal and contextually anchored.

2.2 Traditional Approaches to Emotion Detection

Unimodal emotion identification, the analysis of a single input source, has long been the foundation of emotion detection. These conventional methods frequently concentrate on textual sentiment, verbal cues, or face expressions. Even while these techniques have demonstrated potential in controlled environments, they become less effective in more dynamic, real-world situations. This section examines the fundamental modalities employed in unimodal systems and assesses their advantages and disadvantages.

2.2.1 Unimodal Emotion Recognition Techniques

The analysis of emotional states utilizing a single input modality—such as speech, text, or facial expressions—without including data from other sources is known as unimodal emotion recognition. By concentrating on a single expressive activity, these methods seek to identify emotion, presuming that a single signal can convey sufficient information to deduce the user's emotional state.

Unimodal systems frequently perform poorly in real-world settings where signals may be loud, unclear, or incomplete, despite their effectiveness in specific settings. More resilient multimodal systems have been made possible by this restriction.

Facial Expression Analysis

One of the most extensively researched and used techniques for emotion detection is facial expressions. Early methods, such the Facial Action Coding System (FACS) created by Ekman and Friesen (1978), which divides facial motions into action units (AUs) that correlate to emotional states, depended on geometric features and manually created criteria. Efficient face detection was made possible by machine learning techniques such as the Viola-Jones algorithm, while more modern models automatically extract hierarchical visual features using Convolutional Neural Networks (CNNs).

CNN-based models are used by commercial systems like Affectiva and tools like OpenFace to identify microexpressions in real time. Facial analysis is susceptible to changes in lighting, head posture, and occlusion, though.

Speech and Vocal Tone Analysis

Pitch, tone, intensity, and rhythm are all infused with a variety of emotional indications in speech. Mel-Frequency Cepstral Coefficients (MFCCs), pitch contours, and prosodic features are frequently extracted as part of methods for studying these vocal characteristics. It is now common practice to extract such elements from audio using toolkits like OpenSMILE.

For the purpose of classifying emotions, deep learning models—specifically, Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks—are good at capturing the temporal dynamics of voice data. However, performance can be significantly impacted by background noise, linguistic variances, and speaker variability.

Text-Based Emotion Analysis

Text-based emotion recognition, which relies on extracting syntactic and semantic information from written language, is widely used in sentiment analysis. Previous mod-

els used rule-based systems or bag-of-words characteristics. Transformer-based models, such LSTM or BERT, are presently used in more sophisticated methods to capture more emotional context and complexity.

Emotion lexicons like EmoLex and NRC Emotion Lexicon can be used to link words to emotional categories. These approaches struggle with irony, sarcasm, and cultural context, but they perform well when assessing online content like reviews or tweets.

2.2.2 Limitations of Unimodal Approaches

Despite significant advancements, unimodal systems—which are centered on voice, text, or facial expressions—frequently fall short of capturing the entire complexity of human emotional expression. The presence of noise, occlusions, and ambiguity in real-world settings makes it challenging for a single modality to function consistently. For instance, although verbal emotion detection is extremely sensitive to background noise and speaker fluctuation (Schuller et al., 2011), facial expression analysis may be impacted by head motions, lighting, or occlusions (Zeng et al., 2009). The non-verbal clues required to recognize sarcasm or emotional nuances are also absent from text-based systems. In naturalistic environments where emotional cues are complex, overlapping, or even contradicting, these shortcomings are more noticeable. A unimodal model can misclassify someone who says, ”I’m fine,” with a sorrowful expression and a flat tone. Furthermore, context, which is crucial to the expression and interpretation of emotions, is typically overlooked by unimodal systems (Wieser Brosch, 2012). They consequently have trouble generalizing, particularly across situations, languages, and cultures.

Multimodal emotion recognition, which combines input from several sources for a more comprehensive and precise interpretation of emotional states, is becoming more and more popular in research as a solution to these problems.

2.3 Emergence of Multimodal Emotion Recognition

The shortcomings of unimodal emotion detection systems have sparked the creation of multimodal emotion recognition (MER) techniques, which combine several data sources, including text, audio, and facial expressions, to provide a more accurate and nu-

anced understanding of human emotions. This integration improves the resilience and dependability of emotion detection systems by simulating the human capacity to perceive emotions from several simultaneous signals.

2.3.1 Motivation for Multimodal Approaches

Humans integrate several sensory clues at once rather than depending on a single modality to comprehend the emotions of others. Motivated by this inherent capability, MER systems seek to capture the redundant and complimentary information provided by several modalities. For instance, when a neutral facial expression and a frustrated tone of voice are combined, the algorithm can accurately identify the emotional state.

According to Lian et al. (2023), multimodal systems offer greater accuracy in emotion classification tasks and are more resilient to missing or noisy data in any one modality. Combining text, audio, and visual inputs greatly enhances recognition accuracy in dynamic contexts including video conferences, social media interactions, and conversations, according to their survey on deep learning-based MER.

"Multimodal learning captures complex affective signals more effectively, reducing the ambiguity present in single-channel inputs." — Lian et al., 2023.

According to Khalane et al. (2025), multimodal systems can better utilize contextual cues by coordinating visual and auditory patterns with the semantics of spoken or written content. They contend that the context in which emotions are communicated is crucial.

2.3.2 Fusion Techniques

The fusion strategy, which integrates various modalities, is essential to MER. Three main fusion strategies exist:

- Early Fusion (Feature-Level Fusion): Builds a single model by combining low-level features or raw data from many modalities. It is straightforward and computationally effective, although it may be sensitive to data that is not aligned.
- Late Fusion (Decision-Level Fusion): After processing each modality separately, the outputs are combined to make the final choice (e.g., via voting or averaging). Although it is more adaptable, cross-modal interactions might not be captured.

- Hybrid Fusion: Combining the two approaches with the goal of jointly learning intra- and inter-modal correlations. By dynamically weighing the significance of each modality using attention mechanisms, recent architectures like Multimodal Transformers and Fine-grained Interaction Fusion models (Lian et al., 2023) have demonstrated encouraging outcomes.

Particularly in tasks involving nuanced or complicated emotions, these fusion strategies aid models in becoming more adaptive and context-aware.

2.3.3 Benchmark Datasets

Because benchmark datasets with aligned multimodal data are now available, the development of trustworthy MER systems has advanced considerably:

- IEMOCAP (Interactive Emotional Dyadic Motion Capture): One of the most important and frequently referenced sources for multimodal emotion research is the IEMOCAP dataset. It includes over 12 hours of video footage of professional actors interacting with one another. The dataset consists of:
  - Modalities: Audio, video, and text (transcribed speech).
  - Labels: Categorical emotions such as happiness, anger, sadness, neutral, frustration, and excitement.
  - Features: Facial expressions, body gestures, speech tone, and semantic content.

  IEMOCAP is perfect for training models that need to function in both controlled and natural environments because each session consists of both scripted and unplanned talks. Multimodal systems that examine emotional dynamics in conversations are commonly benchmarked using it.

- CMU-MOSEI (Multimodal Opinion Sentiment and Emotion Intensity): Among the most extensive and varied datasets for multimodal emotion identification is the CMU-MOSEI dataset. More than 23,000 brief YouTube video snippets spanning a variety of subjects, presenters, and emotional tenors are included.

  - Modalities: Audio, video, and text (transcribed speech).
  - Emotion labels: Happiness, sadness, anger, fear, disgust, surprise, and neutral. using tools such as FACET (for facial features) and COVAREP (for vocal features).

  CMU-MOSEI's fine-grained mood and emotion annotations at the utterance level are what make it unique. This enables models to pick up on tiny emotional changes in a variety of speakers' speech patterns, tones, and facial expressions.

- MELD (Multimodal EmotionLines Dataset): Multi-party talks, like those in TV series or meetings, are intended to be reflected by MELD. It is built using conversation from the television show Friends and is an expansion of the EmotionLines dataset.
  - Modalities: Text, audio, and video.
  - Emotion labels: Happiness, sadness, anger, fear, disgust, surprise, and neutral.
  - Structure: Approximately 1,400 dialogues with 13,000+ utterances, each labeled with the speaker's emotion and dialogue context..

  MELD places a strong emphasis on contextual emotion detection, or recognizing how feelings change depending on how a discussion has gone in the past. It is among the rare datasets that takes speaker context and group interaction into account.

With the use of these datasets, multimodal models may be rigorously tested and benchmarked in realistic settings that include background noise, speaker fluctuation, and spontaneous expressions.

2.4 Deep Learning in Emotion Detection

By providing more accuracy, flexibility, and scalability than conventional machine learning techniques, the use of deep learning in emotion identification has greatly improved the area. The multifaceted, multimodal character of emotional expression is difficult for conventional models to portray since they frequently rely on manually constructed elements and simplistic frameworks. Deep learning, on the other hand, is especially useful

for modeling complex emotive states across a variety of input modalities because it allows models to automatically build hierarchical and abstract representations from raw data.

2.4.1 Benefits of Deep Learning in Emotion Recognition

Convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformer-based models are examples of deep learning architectures that have been widely used to solve the difficulties associated with emotion perception. Facial expressions, voice intonations, and linguistic content are just a few examples of the spatial and temporal aspects that these models demonstrate a high ability to learn.

In 2021, Abbaschian, Sierra-Sosa, and Elmaghraby give a thorough review of deep learning applications in voice emotion identification. Their results show that CNN and RNN-based models perform better than traditional approaches by successfully capturing both static and dynamic acoustic characteristics. These models' applicability is further demonstrated by their capacity to generalize in noisy and unstructured contexts.

"Deep learning algorithms offer a substantial improvement over conventional approaches by capturing the complex and nonlinear characteristics of emotional signals" (Abbaschian et al., 2021, p. 7)"

2.4.2 Cross-Modality Applications

Speech Emotion Recognition

Deep learning algorithms examine prosodic and spectral characteristics such pitch fluctuation, energy contours, and Mel-Frequency Cepstral Coefficients (MFCCs) in speech-based emotion identification. In order to capture both local and global temporal dependencies within speech signals, Peng, Wang, and Lu (2021) suggested a multi-scale CNN framework that is connected with attention processes. Their program performed better at classifying emotions on a number of benchmark datasets.

Text-Based Emotion Recognition

The lack of paralinguistic clues makes textual emotion identification a difficult endeavor. However, substantial advancements have been made possible via deep learning.

A hybrid model that combines CNN and Bi-directional Gated Recurrent Units (Bi-GRU) was presented by Bharti et al. (2022) and successfully captures both sequential and spatial patterns in textual data. The feasibility of deep learning for this modality was confirmed by their model, which attained a classification accuracy of above 80% on many datasets.

Facial Expression Recognition

The use of CNNs, which are skilled in identifying spatial hierarchies from face image data, has significantly improved facial emotion identification. According to a review by Saeed, Aslam, and Javaid (2024), CNN-based systems often perform better than conventional algorithms in the recognition of facial affect, especially when trained on a variety of real-world datasets including RAF-DB, AffectNet, and FER2013. These models have demonstrated the ability to identify delicate emotions, such as spontaneous emotional behavior and micro-expressions.

2.4.3 Challenges in Deep Learning-Based Emotion Recognition

Deep learning-based emotion identification systems have a number of drawbacks despite their achievements. One of the main obstacles is the need for sizable, annotated datasets, which are sometimes hard to come by, especially when dealing with multimodal data. Emotional data collection and classification is a laborious and resource-intensive procedure that frequently lacks cultural, linguistic, and contextual variety.

The problem of generality is still another important restriction. Because of variations in recording settings, speaker traits, or emotional expressions across cultures, models trained on one dataset often perform poorly on another.

Data imbalance is still an issue as well. Biased model predictions result from the overrepresentation of emotions like happiness and neutrality and the underrepresentation of emotions like fear, disgust, and surprise.

Lastly, a recurring issue is model interpretability. The majority of deep learning models function as ”black boxes,” providing little information about how they make decisions. Particularly in delicate applications like social robotics, education, or mental health, this lack of transparency might impede adoption and trust.

2.5 Emotion Detection in Virtual Reality (VR)

Although virtual reality (VR) presents special difficulties, it also provides immersive settings that are perfect for researching human emotion. One significant problem that restricts conventional facial expression analysis is face occlusion brought on by head-mounted displays (HMDs). A MobileNetV2-based model was presented by Zhang et al. (2023) and modified for virtual reality. It did well for fundamental emotions like happiness but had trouble with more nuanced displays.

Multimodal techniques are becoming more popular as a way to get around these restrictions. A deep learning architecture that integrates biosignals from the brain, trunk, and peripherals was presented by Indrasiri et al. (2024), who demonstrated enhanced emotion categorization based on arousal and valence levels.

EEG-based techniques have also showed potential in virtual reality. Yu et al. (2024) showed that brain signals are viable for emotion detection in immersive environments by using EEG signals to accurately categorize six fundamental emotions.

2.6 Embodied Conversational Agents (ECAs)

Embodied Conversational Agents (ECAs) are virtual characters that combine verbal and nonverbal communication to mimic in-person human interaction. ECAs, in contrast to conventional chatbots, are outfitted with visual representations, frequently animated avatars, that augment conversation through gestures, facial expressions, and body language.

ECAs have been popular in customer service, healthcare, and education because human-like contact increases user trust and engagement. By modifying their replies according to the user's emotional state, emotion-aware ECAs go one step further and facilitate more organic and sympathetic dialogues. ECAs are particularly useful for creating realistic social simulations and training scenarios in immersive settings like virtual reality.

Despite their potential, developing ECAs that interpret and express emotions accurately remains a challenge, particularly in real-time, multimodal settings.

2.7 Multimodal Emotion Detection for ECAs in VR

Embodied Conversational Agents (ECAs) in Virtual Reality (VR) greatly benefit from multimodal emotion recognition, which integrates language, voice tone, and facial clues to better comprehend users. This is essential in immersive environments where the user experience is influenced by emotional context. ECAs can react more organically by modifying their tone, gestures, and facial expressions in real time by identifying emotions across a variety of channels. Nevertheless, issues including real-time processing, sensor constraints, and the complexity of data fusion continue to impede smooth deployment.

2.7.1 Haptic Feedback and Remote Touch in Emotion Communication

Although visual, auditory, or textual cues are a major part of classic affective computing systems, new developments in remote haptic technology indicate that touch, even when digitally mediated, can greatly influence emotional experience and interpretation. This understudied modality was investigated by Wang and Quek (2010), who created a tactile system that provides emotive touch cues during storytelling sessions.

Using a shape-memory alloy armband, their method coupled audio narration with remote upper-arm pressure to explore if touch could improve emotional experience. According to the findings of their carefully monitored laboratory study, individuals who received contextualized remote touch during a story session reported significantly less melancholy as well as trends toward less negative affect and more joviality.

The idea of ”immediacy” in touch—that is, that affective touch is perceived directly, without encoding or decoding—is a crucial takeaway from their research. However, context is essential to the effectiveness of this immediacy. In order to allow users to naturally create emotional associations, the authors suggest two design guidelines: (1) always pair touch with a secondary communication modality (such as voice) and (2) refrain from giving touch gestures symbolic meaning.

The promise of haptic feedback in embodied conversational beings is demonstrated by these findings, especially in virtual situations where physical co-presence is not present. Contextualized touch has the potential to greatly improve user experience and empathy in

human–agent interaction for ECAs that mimic real-world emotional intelligence.

2.8 Context-Aware Emotion Detection

Agarwal and Gupta's recent research from 2022 highlights how context is becoming more and more crucial for identifying emotions, especially in unrestricted settings. Their research suggests that exterior aspects, including items in the scene, background elements, and spatial arrangements, might have a major impact on emotional perception, whereas traditional techniques mostly concentrate on face expressions or body clues.

Their suggested system uses sophisticated deep learning models, such as a hybrid CNN-RNN architecture (H-CNN), to extract a wide range of visual features from both photos and videos, such as color, shape, texture, and histogram-based descriptors. The model captures not only the subject's look but also the context in which emotions are portrayed by integrating object identification with image caption production.

Agarwal and Gupta's recent research from 2022 highlights how context is becoming more and more crucial for identifying emotions, especially in unrestricted settings. Their research suggests that exterior aspects, including items in the scene, background elements, and spatial arrangements, might have a major impact on emotional perception, whereas traditional techniques mostly concentrate on face expressions or body clues.

Their suggested system uses sophisticated deep learning models, such as a hybrid CNN-RNN architecture (H-CNN), to extract a wide range of visual features from both photos and videos, such as color, shape, texture, and histogram-based descriptors. The model captures not only the subject's look but also the context in which emotions are portrayed by integrating object identification with image caption production.

Chapter 3

Implementation

This chapter covers the datasets, techniques, and methods used in our multimodal emotion detection challenges. Our method was tested using the MELD dataset, which includes text, audio, video, and facial expressions.

3.1 Datasets

The Multimodal EmotionLines Dataset (MELD) has been developed by augmenting and expanding the EmotionLines dataset. MELD has the same conversation situations seen in EmotionLines, but it further incorporates audio and visual modalities with text. MELD comprises almost 1,400 lines and 13,000 utterances from the television sitcom Friends. Numerous speakers engaged in the discussions. Every statement in a discourse has been categorized by one of the following seven emotions: Anger, Disgust, Sadness, Joy, Neutral, Surprise, and Fear. MELD includes sentiment annotation (positive, negative, and neutral) for each speech.

3.1.1 Data Preprocessing

The Multimodal EmotionLines Dataset (MELD), sourced from conversations in the television series Friends, serves as a comprehensive resource for emotion detection, including text, audio, and video modalities annotated with seven emotions: neutral, anger, contempt, fear, joy, sorrow, and surprise. To use this dataset for our multimodal emotion recognition model, we established a comprehensive preprocessing pipeline to ready

the data for the BERT, Wav2Vec2, and Vision Transformer (ViT) models included in our architecture. The pipeline, mostly executed inside the MELDMultimodalDataset class, standardizes text, audio, video (denoted as face in the code), and extracted face modalities, hence assuring compatibility with the model's encoders. Furthermore, we created a facial extraction script to delineate facial areas from unprocessed video files, therefore augmenting the model's capacity to detect emotion-specific visual indicators. This section delineates the preprocessing stages for each modality, emphasizing their purpose and importance in facilitating successful feature extraction and emotion categorization.

3.1.2 Dataset Structure and Assumptions

The MELD dataset is structured into training, validation, and testing subsets, with data included in `.pt` (PyTorch) files located at the directories indicated in `DATA_PATHS`. Each `.pt` file comprises a sample including keys for speech (text), audio (waveform), face (video frame), and emotion (label). Moreover, the retrieved facial data is preserved in distinct `.pt` files located inside `FACE_PATHS`, created by the face extraction software. The preprocessing assumes that:

- Text: Utterances are raw strings of dialogue.
- Audio: Waveforms are sampled at 16 kHz, stored as 1D tensors.
- Video: Single RGB frames (referred to as face in the code) are stored as NumPy arrays of shape [height, width, 3].
- Extracted Face: Preprocessed face images, extracted from raw videos, are stored as tensors of shape [3, 224, 224] in separate .pt files.
- Emotion: Labels are either strings (e.g., "joy") or integers (0–6), mapped using `EMOTION_MAP`.

The preprocessing pipeline utilizes these inputs, converting them into forms suitable for the model while addressing missing or erroneous data to guarantee resilience.

3.2 Text Preprocessing

The text modality encapsulates the vocal content of MELD's conversations, exemplified by phrases like "I'm so excited!" We implemented a number of preprocessing processes to standardize and encode the text data for the BERT-based TextEncoder.

3.2.1 Tokenization with BERT

We used the `bert-base-uncased` tokenizer from the Hugging Face `transformers` library to process each utterance. The tokenizer splits text into subword tokens (e.g., "excited" might become "ex" and "##cited"), converts all characters to lowercase for consistency, and adds special tokens: `[CLS]` at the start (used for classification) and `[SEP]` at the end (marking the sentence boundary). For example, "I'm so excited!" is tokenized into a sequence like `[CLS]`, `I`, `'m`, `so`, `ex`, `##cited`, `!`, `[SEP]`. Lowercasing ensures that variations like "Excited" and "excited" are treated identically, leveraging BERT's pretrained vocabulary.

3.2.2 Truncation and Padding

To maintain consistent input dimensions, we established a maximum sequence length of 128 tokens (`max_text_len=128`). Utterances exceeding 128 tokens are shortened to preserve the first segment, which generally encompasses the essential emotional content. Shorter utterances are supplemented with `[PAD]` tokens to attain a total of 128 tokens. This standardization is crucial for batch processing, since neural networks require fixed-size inputs. An attention mask is created, including 1s for actual tokens and 0s for padding, thereby ensuring that BERT focuses only on significant content.

3.2.3 Encoding to Numerical IDs

Each token is assigned a number ID from BERT's lexicon, resulting in a tensor of `input_ids` with a shape of `[128]`. For example, "excited" may correspond to ID 7560. The attention mask is a tensor with a shape of `[128]`. The tensors are compressed to eliminate singleton dimensions (e.g., from `[1, 128]` to `[128]`) and returned as PyTorch

tensors, prepared for the `TextEncoder`.

3.2.4 Significance

These procedures conform the text data to BERT's specifications, facilitating the model's extraction of contextual characteristics from conversations. The constant length and attention mask facilitate fast batch processing, while lowercasing and subword tokenization use BERT's pre-trained knowledge, making it appropriate for MELD's informal, conversational language. This preprocessing converts unrefined utterances into an organized style that encapsulates semantic subtleties essential for emotions such as pleasure or sarcasm.

3.3 Audio Preprocessing

The auditory modality in MELD offers voice cues, including tone and pitch, which are essential for emotion identification. The function preprocesses the waveforms to ensure compatibility with the Wav2Vec2 model (facebook/wav2vec2-base-960h).

3.3.1 Loading and Validating Audio

Audio data is included in `.pt` files as one-dimensional tensors, presumed to be sampled at 16 kHz. The code retrieves the audio via `sample['audio'].squeeze()`, eliminating any singleton dimensions (e.g., transforming `[1, length]` to `[length]`). To address missing or distorted audio, it verifies if the tensor is empty (`numel() == 0`) or consists entirely of zeros (`torch.all(audio == 0)`). In these instances, a zero tensor of 16,000 samples (1 second at 16 kHz) is used as a placeholder, guaranteeing that the model can handle the sample without errors.

3.3.2 Standardizing Audio Length

The code mandates a fixed length of 80,000 samples (`max_audio_len=80000`), corresponding to 5 seconds at a sampling rate of 16 kHz. Extended audio segments are reduced to the first 80,000 samples, maintaining the original emotional context. Shorter

clips are augmented with zeros to reach 80,000 samples. This standardization guarantees uniformity in audio input lengths, thereby enabling batch processing.

3.3.3 Output Format

The outcome is a tensor with a shape of `[80000]`, signifying the waveform. This tensor is input directly into the `AudioEncoder`, which uses `Wav2Vec2` to extract speech characteristics without requiring other transformations such as spectrograms.

3.3.4 Significance

Standardizing audio length enables efficient batch processing, while the placeholder for invalid audio ensures robustness. The 5-second window captures sufficient vocal cues for emotions like anger or sadness, balancing computational efficiency with contextual richness. By aligning with Wav2Vec2's 16 kHz sampling rate, the preprocessing leverages the model's pre-trained speech representations, enhancing its ability to detect emotional nuances in tone.

3.4 Video Preprocessing

3.4.1 Video Modality

The video modality, referred to as `face` in the code, consists of individual RGB frames extracted from MELD discussions, illustrating visual scenes. These frames are processed by the *VideoEncoder*, which employs ViT (`google/vit-base-patch16-224`) in both models, with an optional ConvNeXt + FRAM pathway used in Model B. Video preprocessing remains mostly consistent across models, with some variations in the application of specific image transformations.

3.4.2 Loading Video Frames

Frames are extracted from `sample['face']` in `.pt` files, which are presumed to be NumPy arrays of shape $[\text{height}, \text{width}, 3]$. The array is first transformed into a PyTorch

tensor, then transposed to a channels-first format $[3, \text{height}, \text{width}]$. It is subsequently normalized from the pixel value range $[0, 255]$ to $[0, 1]$ by dividing by 255.0. Model A specifically resizes frames to $[224, 224]$ prior to normalization, whereas Model B performs direct normalization, with resizing typically handled during the data preparation phase.

3.4.3 Applying Transformations

Transformations are applied via `torchvision.transforms`:

Model A:

Base Transformations: Resize to $[224, 224]$, normalize with ImageNet mean $[0.485, 0.456, 0.406]$ and standard deviation $[0.229, 0.224, 0.225]$.

Training Augmentations (if `use_augmentations=True`):

- RandomHorizontalFlip ($p = 0.5$): Simulates mirrored views.
- RandomRotation ($\pm 20°$): Adds orientation variability.
- ColorJitter (brightness, contrast, saturation: 0.3): Mimics lighting changes.
- RandomAffine (degrees=20, translation=0.15, scale=0.85–1.15): Emulates camera motions.

Model B:

Model B:

Base Transformations: Normalize with ImageNet mean/std $\left[0.485, 0.456, 0.406\right]$ and $\left[0.229, 0.224, 0.225\right]$ (no explicit resize, assuming input is $\left[224,\ 224\right]$).

Training Augmentations: Identical to Model A if enabled. The resize step in Model A ensures compatibility with ViT's input size, while Model B assumes pre-sized frames, a minor difference with negligible impact.

3.4.4 Output Format

The output is a tensor of shape $[\text{batch_size}, 3, 224, 224]$, ready for the VideoEncoder.

3.4.5 Significance

Normalization aligns frames with ViT's pretrained ImageNet statistics, leveraging its visual feature extraction capabilities. Augmentations enhance generalization to MELD's dynamic environments (e.g., varying lighting, angles), improving robustness for body language and contextual cues. The shared video preprocessing ensures both models extract comparable visual features, though Model B's optional ConvNeXt+FRAM may enhance feature quality (61.12%,).

3.5 Extracted Face Preprocessing

The face modality in MELD provides critical visual cues for emotion detection, capturing facial expressions (e.g., smiles for joy, furrowed brows for anger). Preprocessing adapts face data for the FaceEncoder (VideoMAE for Model A, ConvNeXt+FRAM for Model B), handling variability in data quality (e.g., occlusion, illumination). Model A processes frame sequences with Action Units (AUs), while Model B uses single frames without AUs, significantly impacting performance.

3.5.1 Loading and Validating Face Data

The facial data in MELD is contained in `.pt` files, with primary data (text, audio, face) located in folders designated by `DATA_PATHS`, and face-specific data (extracted facial frames and AUs) in directories designated by `FACE_PATHS`. The preprocessing starts with the loading and validation of these inputs to confirm their use by the model.

3.5.2 Loading Main Face Data

The `face` field in the primary `.pt` file is presumed to be a NumPy array denoting a single RGB picture (e.g., shape $[\text{height}, \text{width}, 3]$), presumably a sample frame from the discussion. The code retrieves this using `sample.get('face', np.zeros((224, 224, 3)))`, supplying a default zero array if the `face` key is absent. The array is transformed into a PyTorch tensor, transposed to the channels-first format ($[3, \text{height}, \text{width}]$), then normalized to the range $[0, 1]$ by dividing by 255.0. The fallback mechanism guaran-

tees that samples without a primary facial picture may nevertheless undergo processing by using other modalities or the extracted facial data.

3.5.3 Loading Extracted Face Data

Loading and Validating Face Data

Face data is stored in `.pt` files, with main data in `DATA_PATHS` and face-specific data in `FACE_PATHS`. Preprocessing validates inputs to ensure compatibility with the encoders.

Model A (ViT/VideoMAE with Action Units)

- **Main Face Data:** Retrieved from `sample['face']` as a NumPy array $[height, width, 3]$. Converted to a PyTorch tensor, transposed to $[3, height, width]$, and normalized to $[0, 1]$ by dividing by 255.0. A zero tensor $[3, 224, 224]$ is used if missing.
- **Extracted Face Data:** Loaded from face-specific `.pt` files via `torch.load(face_path)`, retrieving the frame key (expected shape $[num_frames, 3, 224, 224]$ or $[3, 224, 224]$). Valid frames set `has_face=True`; invalid or missing frames use a zero tensor $[8, 3, 224, 224]$ with `has_face=False`.
- **Action Units (AUs):** Extracted from the `aus` key as a 12D float tensor, defaulting to zeros $[12]$ if missing. AUs represent facial muscle movements (e.g., AU1: inner brow raise).

Model B (ConvNeXtEnhanced)

- **Main Face Data:** Identical to Model A, yielding $[3, height, width]$ after normalization.
- **Extracted Face Data:** Loaded similarly, but expects a single frame $[3, 224, 224]$. Valid frames set `has_face=True`; invalid/missing frames use a zero tensor $[3, 224, 224]$ with `has_face=False`.
- **Action Units (AUs):** Not extracted, simplifying the pipeline.

3.5.4 Significance

Robust validation ensures both models handle MELD's variability (e.g., missing faces due to detection failures). Model A's sequence and AU processing capture temporal and fine-grained cues but risk noise from padding. Model B's single-frame approach reduces noise, contributing to higher accuracy (61.12%).

Standardizing Face Inputs

Model A (ViT/VideoMAE with Action Units)

- **Frame Sequence Adjustment:** Sequences are standardized to `max_frames` = 8. If `num_frames` ¿ 8, frames are linearly sampled (e.g., indices 0, 2, 4, ..., 14 for 16 frames). If `num_frames` ¡ 8, zero tensors are appended to reach 8 frames. Single frames are unsqueezed to $[1, 3, 224, 224]$ and padded. The output shape is $[8, 3, 224, 224]$.
- **Ensuring Correct Format:** Frames are converted to floats and normalized to $[0, 1]$ by dividing by 255.0, aligning with main face data.
- **Significance:** Standardizing to 8 frames balances temporal information and computational cost, enabling VideoMAE to capture dynamic expressions (e.g., smile formation). Padding introduces noise, reducing efficacy when face data is sparse.

Model B (ConvNeXtEnhanced)

- **Frame Adjustment:** Processes a single frame, requiring no sequence standardization. The frame is validated as $[3, 224, 224]$.
- **Ensuring Correct Format:** Identical to Model A, normalized to $[0, 1]$.
- **Significance:** Single-frame processing avoids padding noise, enhancing robustness and aligning with ConvNeXt+FRAM's focus on spatial features, contributing to higher accuracy (61.12%).

Image Transformations

Both models apply `torchvision.transforms` to face data, ensuring compatibility with their encoders.

Model A (ViT/VideoMAE with Action Units)

- **Base Transformations:**
  - Resize: Frames resized to $[224, 224]$.
  - Normalize: Uses ImageNet mean $[0.485, 0.456, 0.406]$, std $[0.229, 0.224, 0.225]$.
- **Training Augmentations** (if `use_augmentations=True`): Applied per frame, identical to video augmentations (flips, rotations, color jitter, affine transforms).
- **Significance:** Per-frame transformations enhance robustness to variability, but repeated padding may dilute augmentation benefits.

Model B (ConvNeXtEnhanced)

- **Base Transformations:** Normalize with ImageNet mean/std, assuming $[224, 224]$ input (no resize).
- **Training Augmentations:** Identical to Model A.
- **Significance:** Augmentations improve generalization, while single-frame processing ensures clean input for ConvNeXt+FRAM's region-based attention.

Output Format

- **Model A:** Frame sequence $[batch_size, 8, 3, 224, 224]$, AUs $[batch_size, 12]$, `has_face` $[batch_size]$.
- **Model B:** Single frame $[batch_size, 3, 224, 224]$, `has_face` $[batch_size]$.

Overall Significance

Model A's sequence and AU processing aim to capture temporal and fine-grained cues but introduce noise from padding, impacting accuracy (39.83%). Model B's single-frame approach minimizes noise, leveraging ConvNeXt+FRAM's spatial focus for better performance (61.12%, Table 4.8). Robust handling of missing data via `has_face` ensures both models process incomplete inputs, though Model A's complexity amplifies noise effects (Section 5.1).

3.5.5 Dataset and Dataloading

Both models use `MELDMultimodalDataset` to load and preprocess MELD data, but differ in dataloading parameters.

Data Loading

.pt files are loaded from `DATA_PATHS` (main data: text, audio, face) and `FACE_PATHS` (extracted face data). The dataset aligns files by filename, setting `None` for missing face files. `MELDMultimodalDataset` applies modality-specific preprocessing, ensuring consistent tensor formats.

Batching and Collation

A custom `collate_fn` stacks tensors into batches:

- **Text:** Input IDs, attention masks $[\text{batch_size}, 128]$.
- **Audio:** Waveforms $[\text{batch_size}, 80000]$.
- **Video:** Frames $[\text{batch_size}, 3, 224, 224]$.
- **Face:** Sequences $[\text{batch_size}, 8, 3, 224, 224]$ (Model A) or single frames $[\text{batch_size}, 3, 224, 224]$ (Model B).
- **AUs:** $[\text{batch_size}, 12]$ (Model A only).
- **has_face:** $[\text{batch_size}]$.
- **Emotion labels:** $[\text{batch_size}]$.

Dataloading Parameters

- **Model A:** Batch size of 16, no weighted sampling, optional augmentations. Uses 2 worker threads for efficiency.
- **Model B:** Batch size of 8, optional weighted random sampling (inverse class frequency weights), optional augmentations. Also uses 2 worker threads.

Significance: Model B's smaller batch size and weighted sampling address MELD's class imbalance, enhancing training stability and generalization (61.12% vs. 39.83%). Model A's larger batch size may reduce stability on imbalanced data.

Output Format

Batched tensors are ready for model inputs, with shapes as specified per modality.

Significance

The shared dataset structure ensures consistent data access, while Model B's dataloading enhancements likely contribute to its superior performance. Weighted sampling mitigates MELD's imbalanced emotion distribution, a key factor in achieving 61.12% accuracy.

3.6 Model Architecture

This section introduces two variations of a multimodal emotion recognition model designed to classify seven emotions—*neutral*, *anger*, *disgust*, *fear*, *joy*, *sorrow*, and *surprise*—using the MELD dataset. Both models incorporate four modalities: text, audio, video, and facial data.

Each model is instantiated from the `MultimodalEmotionDetector` class and employs pretrained encoders for feature extraction, a cross-attention-based fusion mechanism with dynamic gating, and classification heads supported by auxiliary supervision.

Model A utilizes BERT for text, Wav2Vec2 for audio, the Vision Transformer (ViT) for video analysis, and VideoMAE for facial sequence interpretation, including Action Unit (AU) classification to enhance expression-level detail.

Model B, while using the same text and audio encoders, replaces the video encoder with either ViT or ConvNeXt integrated with an Enhanced Facial Region Attention Module (FRAM), and uses ConvNeXt+FRAM for facial analysis. Unlike Model A, Model B excludes AU classification for simplicity and training efficiency.

Model B demonstrates improved precision, attributed to its advanced visual encoders and refined training methodologies. This section presents a comparative breakdown of the shared and distinct components of both models, including their encoders, fusion mechanisms, classification strategies, and training procedures. Visual diagrams are provided to illustrate architectural designs and data flows for clarity.

3.6.1 Overview of the Multimodal Architecture

Both approaches adopt a modular architecture comprising four modality-specific encoders—`TextEncoder`, `AudioEncoder`, `VideoEncoder`, and `FaceEncoder`—a fusion module (`MultimodalFusion`), and classification heads. Each encoder extracts modality-specific feature embeddings, which are subsequently integrated to capture inter-modal interactions.

Pretrained models are employed to ensure effective feature extraction: both models use BERT for text and Wav2Vec2 for audio. For video and facial processing, **Model A** utilizes the Vision Transformer (ViT) for video and VideoMAE for facial sequences. In contrast, **Model B** employs either ViT or ConvNeXt combined with the Enhanced Facial Region Attention Module (FRAM) for video encoding, and ConvNeXt+FRAM exclusively for facial analysis, enabling more advanced visual understanding.

Both models use a cross-attention mechanism within the fusion module to model inter-modal relationships, along with dynamic gating to adaptively weight the contribution of each modality. While Model A integrates Action Unit (AU) classification for detailed facial expression analysis, Model B streamlines its architecture by focusing solely on emotion classification, improving training efficiency.

To accommodate the challenges posed by MELD's multimodal nature, both models implement a `has_face` flag, which ensures robustness against missing or low-quality facial data. The combination of sophisticated visual encoders and efficient training strategies

enables Model B to achieve higher accuracy in emotion recognition tasks.

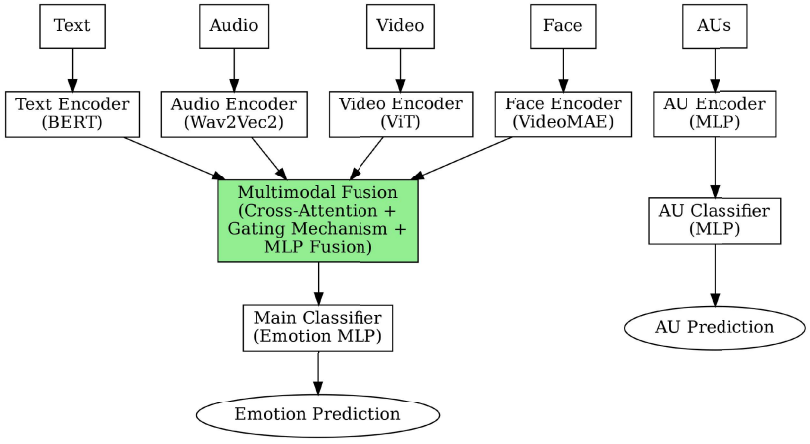

Figure 3.1: Architecture of Model A

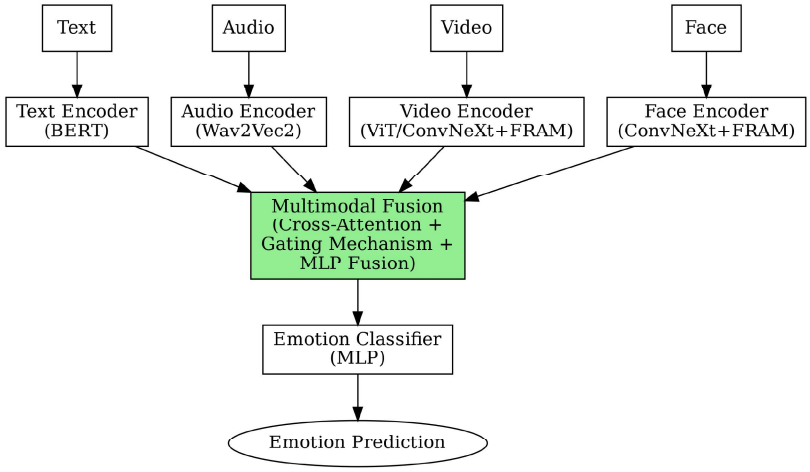

Figure 3.2: Architecture of Model B

3.6.2 Shared Modality Encoders

The text and audio encoders in both models are identical, utilizing pretrained models to ensure effective and robust feature extraction. The following subsection outlines the common components shared between the two architectures.

Text Encoder

The `TextEncoder` in both models processes tokenized text inputs—specifically, input IDs and attention masks with shape $[\text{batch_size}, 128]$—using the `bert-base-uncased`

model from Hugging Face's Transformers library. BERT generates contextualized embeddings for each token, and the final hidden state of the `[CLS]` token (of dimension 768) is used as a summary representation of the entire utterance.

This 768-dimensional embedding is projected to a 256-dimensional vector through a linear transformation layer: `nn.Linear(768, 256)`, aligning it with the model's unified embedding dimension.

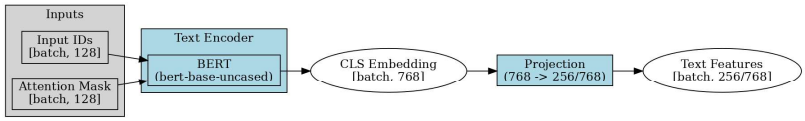

Figure 3.3: bert-base-uncased Text Encoder

To better handle the emotionally nuanced conversations in the MELD dataset, the initial layers of BERT are frozen to retain general language understanding, while the final two encoder layers are fine-tuned to capture dialogue-specific emotional cues. This selective fine-tuning strategy allows the encoder to quickly adapt to the task without overfitting, enhancing its effectiveness across both models.

AudioEncoder

The `AudioEncoder` processes raw audio waveforms with a shape of $[\text{batch_size}, 80000]$, corresponding to 5 seconds of audio sampled at 16 kHz, using the pretrained model:

```
facebook/wav2vec2-base-960h
```

Wav2Vec2, trained on 960 hours of unlabeled speech data, generates a sequence of 768-dimensional embeddings via its feature extractor and transformer encoder.

To obtain a fixed-size representation, adaptive average pooling is applied:

```
nn.AdaptiveAvgPool1d(1)
```

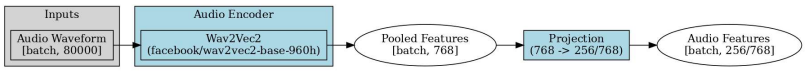

Figure 3.4: wav2vec2-base-960h Audio Encoder

This reduces the variable-length sequence to a single 768-dimensional vector, which is then projected to 256 dimensions using:

```
nn.Linear(768, 256)
```

To preserve general acoustic features, the lower layers of Wav2Vec2 are frozen, while only the final two transformer layers are fine-tuned to capture emotion-specific cues such as pitch, tone, and intonation. This approach ensures expressive and robust audio feature extraction in both models, enhancing performance on MELD's emotionally rich conversations.

3.6.3 Video Encoder

The video encoders vary considerably between the two versions, affecting their visual feature extraction capability.

Model A: Video Encoder

Model A's `VideoEncoder` processes static facial images with the shape:

$$[\text{batch_size}, 3, 224, 224]$$

using the pretrained Vision Transformer model google/vit-base-patch16-224. Each image is divided into fixed-size patches, which are linearly embedded and passed through transformer layers. A special `[CLS]` token is prepended to serve as a global representation of the input.

The final hidden state of the `[CLS]` token, which has a dimensionality of 768, is projected to 256 dimensions using a linear transformation:

```
nn.Linear(768, 256)
```

This 256-dimensional output serves as the visual embedding for emotion classification.

To retain general visual features acquired from ImageNet pretraining while adapting to MELD-specific emotional nuances, only the last two transformer layers are fine-tuned,

Figure 3.5: Video Encoder of Model A

with the earlier layers kept frozen. Although this architecture is efficient and benefits from transfer learning, it may struggle to capture subtle and fine-grained facial details that are essential for accurate emotion recognition.

Model B: Video Encoder

Model B's `VideoEncoder` follows a dual-path strategy, governed by the `use_convnext` parameter. When `use_convnext=False`, the setup mirrors Model A, using ViT with the same configuration. However, when `use_convnext=True`, the encoder integrates ViT with a `ConvNeXtEnhancedFRAM` module based on the `convnext_small` architecture.

The Enhanced FRAM employs six region-specific attention maps (7×7 priors) and channel attention with a reduction ratio of 8. It uses convolutional layers with batch normalization and ReLU activation to generate refined 768-dimensional features. These are concatenated with the 768-dimensional ViT output, resulting in a combined 1536-dimensional feature vector.

This vector is passed through a multilayer perceptron with dropout:

`nn.Linear(1536, 512)` $\rightarrow$ `nn.Linear(512, 256)`, with dropout $= 0.3$

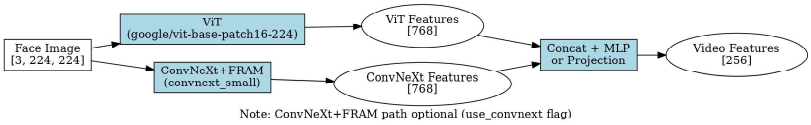

Figure 3.6: Video Encoder of Model B

For fine-tuning, only the last two layers of ViT and stages 2–3 of ConvNeXt are updated, enabling the model to focus on high-level emotional cues. This advanced design significantly improves Model B's accuracy by capturing both fine-grained spatial patterns and broader contextual signals from facial images.

3.6.4 Face Encoder

The face encoders signify a crucial distinction, since Model B's configuration is tailored for single-frame processing.

Model A: Facial Encoder

Model A's `FaceEncoder` processes face frame sequences with the shape:

$$[\text{batch_size},\ 8,\ 3,\ 224,\ 224]$$

using the pretrained MCG-NJU/videomae-base model. VideoMAE is designed to capture both spatial and temporal patterns from video inputs. It uses the final hidden state of the `[CLS]` token to represent the sequence, which is then projected from 768 to 256 dimensions using a linear layer:

```
nn.Linear(768, 256)
```

The model supports a maximum of 8 frames per sample. When fewer frames are available, zero-padding is applied, which may introduce noise and reduce feature quality. To balance adaptation and generalization, only the last two transformer layers of VideoMAE are fine-tuned to learn MELD's facial emotion dynamics.

Additionally, an Action Unit (AU) classifier:

```
nn.Linear(256, 12)
```

is used to predict 12 facial AUs (e.g., AU1 for inner brow raiser), enabling detailed expression analysis. While this auxiliary task adds value for interpretability, it also introduces additional training complexity.

Figure 3.7: Face Encoder of Model A

Model B: Facial Encoder

Model B's `FaceEncoder` processes single facial images of shape $[\text{batch_size}, 3, 224, 224]$ using the `ConvNeXtEnhancedFRAM` module based on `convnext_small`. The Enhanced FRAM applies six region-specific attention maps and channel attention to focus on emotionally salient facial regions.

The resulting 768-dimensional feature vector is projected to 256 dimensions via `nn.Linear(768, 256)`. During training, the initial stages of ConvNeXt are frozen, while stages 2 and 3 are fine-tuned to capture MELD-specific facial expression patterns.

The use of the `has_face` flag ensures robustness by masking facial features when no valid data is available. Unlike Model A, Model B does not include AU classification, simplifying the learning objective. By focusing on static frames—more aligned with MELD's facial data characteristics—Model B achieves higher accuracy with reduced training complexity.

Figure 3.8: Face Encoder of Model B

3.6.5 Multimodal Fusion Mechanism

Both approaches use a comparable MultimodalFusion module, with some variations in implementation.

Model A: Multimodal Integration

The fusion module in Model A integrates 256-dimensional embeddings from the text, audio, video, and facial encoders using 12 pairwise cross-attention mechanisms (e.g., text-to-audio, face-to-video). Each `CrossAttention` layer computes query, key, and value projections via `nn.Linear(256, 256)` and applies scaled dot-product attention to generate attended features.

The attended and original features are concatenated to form a 1024-dimensional vector, which is passed through a linear transformation (`nn.Linear(1024, 1)`) to pro-

duce gating logits. These logits are normalized with a `softmax` function to compute dynamic weights for each modality pair.

Gated features are then aggregated and normalized using `nn.LayerNorm(256)`, followed by a multi-layer perceptron (`nn.Linear(256, 256)` with GELU activation) and a residual connection to retain input information. To ensure robustness, the `has_face` flag is used to nullify facial features when they are absent, preventing noise from affecting the fused representation.

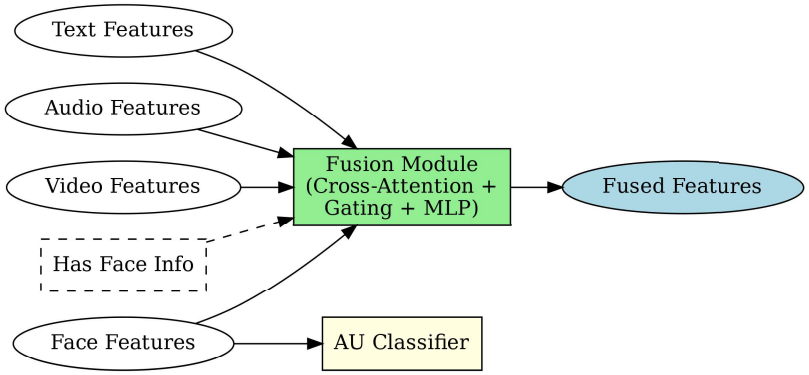

Figure 3.9: Multimodal Integration of Model A

Model B: Multimodal Integration

Model B's fusion module mirrors the design of Model A, employing the same 12 cross-attention modules, gating logic, and transformation architecture. However, unlike Model A, which processes heterogeneous embeddings (768 for text/audio/video and 256 for face), Model B standardizes all modality embeddings to 256 dimensions prior to fusion.

This homogeneity in embedding size reduces the need for dimensional projection and may minimize information loss, contributing to Model B's performance gains. The `has_face` flag is used in the same way as in Model A, ensuring stable operation even when facial data is missing.

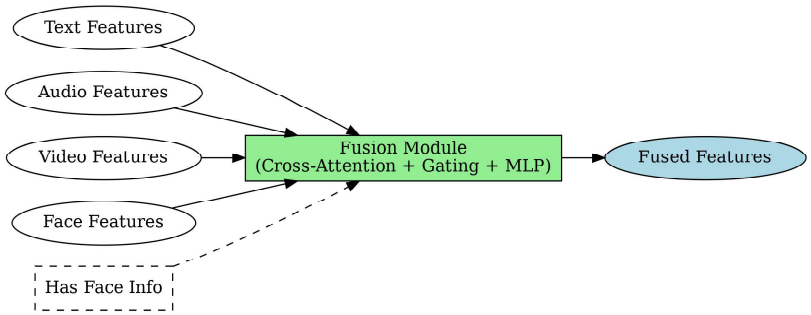

Figure 3.10: Multimodal Integration of Model B

3.6.6 Classification Heads

The categorization heads vary in their supplementary components.

Model A: Classification Mechanisms

The principal classifier in Model A is a two-layer multilayer perceptron (MLP) defined as:

`nn.Linear(256, 128)` → `nn.LayerNorm(128)` → `nn.GELU()` → `nn.Dropout(0.1)` → `nn.Linear(128, 7)`

This classifier predicts emotion logits from the fused multimodal embedding.

Model A also includes auxiliary classifiers for each modality—text, audio, video, and face—each implemented as `nn.Linear(256, 7)`. Additionally, an Action Unit (AU) classifier is used: `nn.Linear(256, 12)`, providing fine-grained facial expression analysis.

Loss contributions are weighted as follows: 0.6 for the primary classifier, 0.1 each for the text, audio, and video auxiliary classifiers, 0.05 for the face classifier (scaled by `has_face`), and 0.05 for the AU classifier. While the AU classifier increases architectural complexity, it enhances facial interpretability by supervising action-unit-specific dynamics.

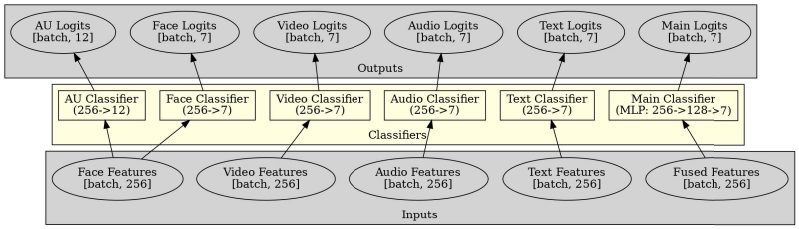

Figure 3.11: Classification Mechanism of Model A

Model B: Classification Mechanisms

Model B adopts the same MLP architecture for its principal classifier as Model A. All auxiliary classifiers for text, audio, video, and facial recognition are implemented as `nn.Linear(256, 7)`.

Unlike Model A, Model B omits the AU classifier to reduce complexity. The loss weights are distributed as follows: 0.6 for the main classifier and 0.1 for each of the four auxiliary classifiers, with the face classifier's contribution scaled based on the presence of valid facial data (`has_face`).

This simplified design eliminates the need to rely on potentially noisy AU labels and allows the model to concentrate more effectively on the emotion recognition task, potentially improving overall accuracy and training efficiency.

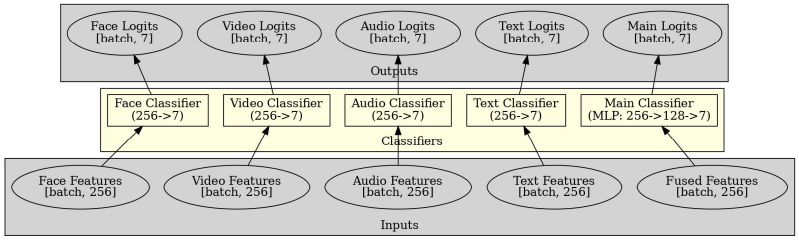

Figure 3.12: Classification Mechanism of Model B

3.6.7 Training Considerations

Model A: Training Considerations

Model A uses cross-entropy loss for emotion classification and `BCEWithLogitsLoss` for Action Unit (AU) prediction. The optimizer is `AdamW` with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-5}. A learning rate scheduler (`StepLR`) is applied with

`step_size=2` and `gamma=0.9` to gradually reduce the learning rate during training.

To preserve general features learned during pretraining, the lower layers of all encoders are frozen, and only the last two layers are fine-tuned. The model is trained for 5 epochs with a batch size of 16, without gradient accumulation.

Checkpoints are saved at the end of each epoch, with the best model selected based on validation accuracy. Although AU supervision provides finer granularity in facial expression analysis, the absence of optimization enhancements—such as gradient accumulation or weighted sampling—may limit Model A's overall performance.

Model B: Training Considerations

Model B focuses solely on emotion classification using cross-entropy loss and omits AU supervision to streamline optimization. It also utilizes the `AdamW` optimizer with the same learning rate (1×10^{-4}) and weight decay (1×10^{-5}), along with the `StepLR` scheduler.

To address memory constraints, Model B incorporates gradient accumulation (`accumulation_`
effectively increasing the batch size to 8 and stabilizing training dynamics. Optional weighted sampling is used to compensate for class imbalance in MELD, ensuring that underrepresented emotions receive adequate learning signals.

Additionally, Model B applies data augmentation techniques—such as random horizontal flips, rotations, color jitter, and affine transformations—to improve the robustness of face representations. Selective fine-tuning is applied to later layers of the ConvNeXt backbone (specifically stages 2 and 3), allowing for dataset-specific adaptation.

Model B is trained over 20 epochs, with model checkpoints saved after each epoch. The integration of gradient accumulation, weighted sampling, and data augmentation contributes to improved generalization and likely enhances overall accuracy.

3.6.8 Significance of the Architecture

Both designs are modular and robust, leveraging pretrained encoders for comprehensive feature extraction and cross-attention mechanisms for effective inter-modal interaction. The inclusion of Action Unit (AU) classification in Model A enhances facial

expression analysis but introduces additional complexity. This may negatively impact performance due to potential inaccuracies in AU annotations or noise from sequence-based facial processing, such as zero-padding.

Model B, by contrast, integrates `ConvNeXt+FRAM` for video encoding and employs single-frame facial processing. Coupled with a simplified loss function, this design aligns more closely with the MELD dataset's characteristics, reducing noise and improving accuracy. Furthermore, Model B benefits from advanced training strategies, including gradient accumulation and weighted sampling, which significantly enhance convergence stability and generalization.

Both models incorporate the `has_face` flag to manage missing facial data; however, Model B's architecture is particularly well-suited to MELD's natural conversational settings. As a result, it offers superior efficacy in multimodal emotion recognition tasks.

Chapter 4

Experiments and Results

The experimental procedures, findings, and results of the multimodal emotion detection system described in Chapter 3 are delineated in this chapter. The model was assessed in a variety of configurations, including text-only, text+audio, audio+video, audio+video+text, and audio+video+text+face, using the MELD dataset.

4.1 Text-Only Modality

4.1.1 Test Set Evaluation Results

The performance of the proposed multimodal emotion detection model was evaluated on the test set. The results include the overall test loss, accuracy, and a detailed classification report indicating the precision, recall, and F1-score for each emotion category.

- **Test Loss:** 1.0597
- **Test Accuracy:** 63.57%

Class	Precision	Recall	F1-Score	Support
Neutral	0.71	0.88	0.78	1256
Anger	0.55	0.35	0.43	345
Disgust	0.75	0.09	0.16	68
Fear	0.25	0.10	0.14	50
Joy	0.54	0.55	0.54	402
Sadness	0.41	0.21	0.28	208
Surprise	0.56	0.56	0.56	281
Accuracy	63.57% (Total samples: 2610)			
Macro Avg	0.54	0.39	0.41	2610
Weighted Avg	0.61	0.63	0.61	2610

Table 4.1: Text Only Classification Report

4.2 Text + Audio Modalities

4.2.1 Fusion Model Test Results

The following table summarizes the test performance of the fusion model and its constituent modalities—text and audio. The model was evaluated using standard metrics such as loss and accuracy.

- **Test Loss:** 1.0584
- **Fusion Accuracy:** 0.6342
- **Text Accuracy:** 0.6349
- **Audio Accuracy:** 0.4828

Emotion Class	Precision	Recall	F1-Score	Support
Neutral	0.71	0.87	0.78	1256
Anger	0.52	0.38	0.44	345
Disgust	0.60	0.09	0.15	68
Fear	0.31	0.08	0.13	50
Joy	0.54	0.55	0.55	402
Sadness	0.46	0.24	0.32	208
Surprise	0.54	0.54	0.54	281
Overall Accuracy	0.6342			
Macro Avg	0.53	0.39	0.41	2610
Weighted Avg	0.61	0.63	0.61	2610

Table 4.2: Text + Audio Classification Report

4.3 Text + Video Modalities

4.3.1 Fusion Model Test Results

The following table summarizes the test performance of the fusion model and its constituent modalities—text and video. The model was evaluated using standard metrics such as loss and accuracy.

- **Test Loss:** 1.0612
- **Fusion Accuracy:** 0.6319
- **Text Accuracy:** 0.6342
- **Video Accuracy:** 0.4824

Emotion Class	**Precision**	**Recall**	**F1-Score**	**Support**
Neutral	0.70	0.88	0.78	1256
Anger	0.54	0.37	0.44	345
Disgust	1.00	0.06	0.11	68
Fear	0.38	0.06	0.10	50
Joy	0.56	0.52	0.54	402
Sadness	0.48	0.15	0.23	208
Surprise	0.49	0.59	0.53	281
Overall Accuracy		0.6319		
Macro Avg	0.59	0.38	0.39	2610
Weighted Avg	0.62	0.63	0.60	2610

Table 4.3: Text+Video Classification Report

4.4 Audio + Video Modalities

4.4.1 Fusion Model Test Results

The following table summarizes the test performance of the fusion model and its constituent audio and video modalities. The model was evaluated using standard metrics such as loss and accuracy.

- **Test Loss:** 1.4934

- **Fusion Accuracy:** 0.4836
- **Audio Accuracy:** 0.4824
- **Video Accuracy:** 0.4824

Emotion Class	Precision	Recall	F1-Score	Support
Neutral	0.70	0.88	0.78	1256
Anger	0.53	0.38	0.45	345
Disgust	1.00	0.06	0.11	68
Fear	0.36	0.08	0.13	50
Joy	0.54	0.55	0.54	402
Sadness	0.50	0.17	0.26	208
Surprise	0.55	0.54	0.54	281
Overall Accuracy		0.6365		
Macro Avg	0.60	0.38	0.40	2610
Weighted Avg	0.62	0.64	0.60	2610

Table 4.4: Audio+Video Classification Report

4.5 Text + Audio + Video Modalities

4.5.1 Fusion Model Test Results

The following table summarizes the test performance of the fusion model and its constituent modalities—text, audio, and video. The model was evaluated using standard metrics such as loss and accuracy.

- **Test Loss:** 1.0544
- **Fusion Accuracy:** 0.6365
- **Text Accuracy:** 0.6323
- **Audio Accuracy:** 0.4824
- **Video Accuracy:** 0.4824

Emotion Class	Precision	Recall	F1-Score	Support
Neutral	0.70	0.88	0.78	1256
Anger	0.53	0.38	0.45	345
Disgust	1.00	0.06	0.11	68
Fear	0.36	0.08	0.13	50
Joy	0.54	0.55	0.54	402
Sadness	0.50	0.17	0.26	208
Surprise	0.55	0.54	0.54	281
Overall Accuracy		0.6365		
Macro Avg	0.60	0.38	0.40	2610
Weighted Avg	0.62	0.64	0.60	2610

Table 4.5: Text+Audio+Video Classification Report

4.6 Text + Audio + Video + face Modalities with ViT/VideoMAE

4.6.1 Fusion Model Test Results

The following section summarizes the test performance of the fusion model when using only non-text modalities—audio, video, and face. The evaluation includes the overall test loss and accuracy across all individual and fused modalities.

- **Test Loss:** 2.6585
- **Fusion Accuracy:** 0.3983
- **Text Accuracy:** 0.4813
- **Audio Accuracy:** 0.4901
- **Video Accuracy:** 0.3884
- **Face Accuracy:** 0.4130

Emotion Class	Precision	Recall	F1-Score	Support
Neutral	0.49	0.73	0.59	1256
Anger	0.28	0.08	0.13	345
Disgust	0.07	0.01	0.02	68
Fear	0.00	0.00	0.00	50
Joy	0.17	0.12	0.14	402
Sadness	0.08	0.02	0.03	208
Surprise	0.11	0.11	0.11	281
Overall Accuracy		0.3983		
Macro Avg	0.17	0.15	0.15	2610
Weighted Avg	0.32	0.40	0.34	2610

Table 4.6: Text+Audio+Video+Face Classification Report with ViT/VideoMAE

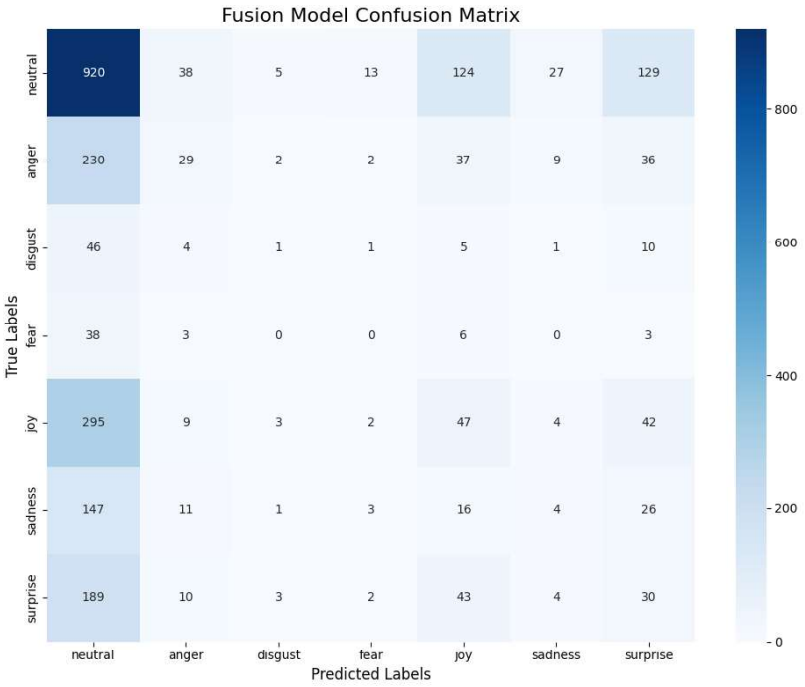

Figure 4.1: Text + Audio + Video Modalities Confusion Matrix

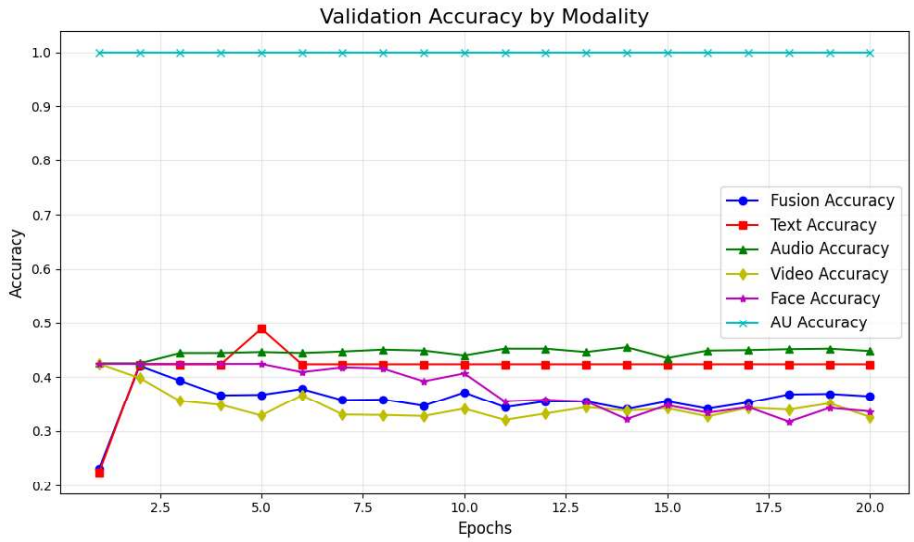

Figure 4.2: Text+Audio+Video+Face with Action unit model validation Accuracy Graph

4.7 Text + Audio + Video + face Modalities with Convanxt

4.7.1 Fusion Model Test Results

The following results summarize the performance of the fusion model on the test set, including loss and accuracy across the fusion and individual modalities.

- **Test Loss:** 1.2533
- **Fusion Accuracy:** 0.6112
- **Text Accuracy:** 0.6258
- **Audio Accuracy:** 0.4813
- **Video Accuracy:** 0.4717
- **Face Accuracy:** 0.4815

Emotion Class	Precision	Recall	F1-score	Support
Neutral	0.69	0.86	0.77	1256
Anger	0.54	0.37	0.44	345
Disgust	0.20	0.09	0.12	68
Fear	0.20	0.06	0.09	50
Joy	0.53	0.47	0.50	402
Sadness	0.37	0.15	0.22	208
Surprise	0.48	0.57	0.53	281
Overall Accuracy		0.6112		
Macro Avg	0.43	0.37	0.38	2610
Weighted Avg	0.58	0.61	0.58	2610

Table 4.7: Text+Audio+Video+face Modalities with ConvNeXt

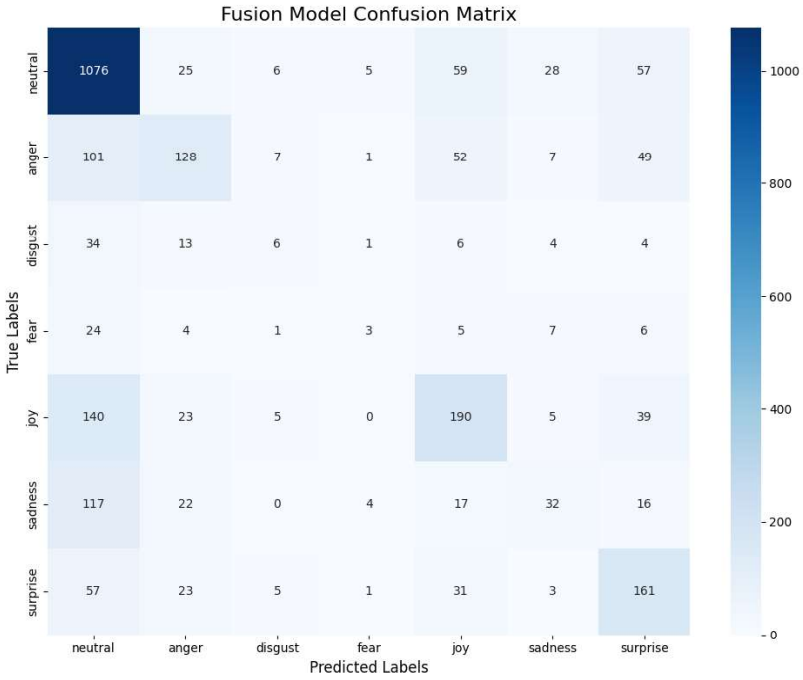

Figure 4.3: Enter Caption

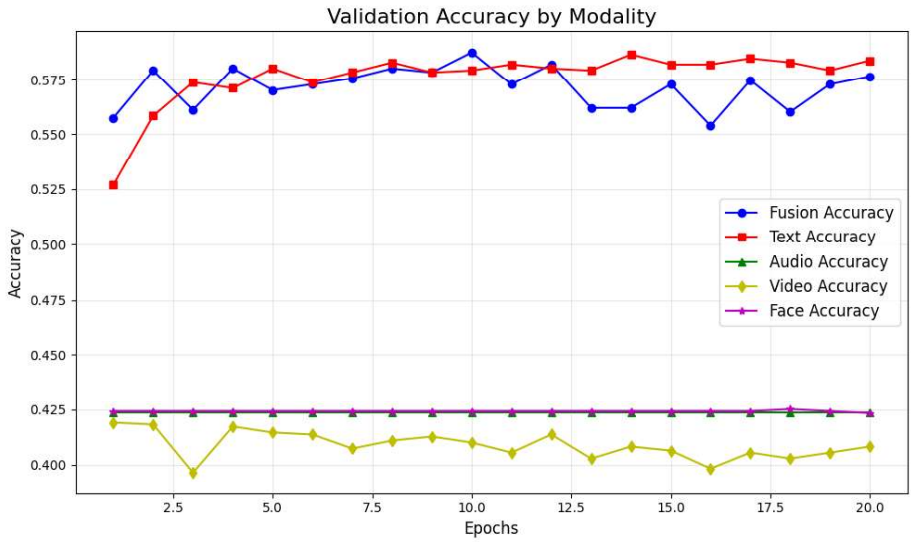

Figure 4.4: Text+Audio+Video+Face with convanxt validation Accuracy Graph

4.8 Comparative Analysis of ConvNeXtEnhancedFRAM and ViT/VideoMAE Architectures

4.8.1 Overview

This section contrasts two multimodal emotion recognition models: one using the `ConvNeXtEnhancedFRAM` architecture and the other employing `ViT`/`VideoMAE` architectures. The `ConvNeXtEnhancedFRAM` model routinely surpasses its `ViT`/`VideoMAE` counterpart owing to its superior spatial feature extraction, improved alignment with the MELD dataset features, and greater resilience to missing or inconsistent data.

4.8.2 Architectural Benefits

Hierarchical Feature Extraction: `ConvNeXt` employs a convolutional, hierarchical architecture that captures both detailed and overarching characteristics. This design is particularly suitable for emotion recognition tasks, where nuanced spatial details are essential. In contrast, `ViT` analyzes flattened patches, potentially sacrificing local texture details—especially for smaller datasets such as MELD.

The `FRAM` module utilizes learnable region-specific and channel-specific attention to highlight significant face areas, including the eyes and lips. This specialized approach

enhances the identification of subtle emotional signals, whereas the global attention used in ViT and VideoMAE applies uniform treatment to all areas, potentially reducing focus on key regions.

Hybrid Feature Integration: ConvNeXtEnhancedFRAM may potentially amalgamate ConvNeXt and ViT features, thereby effectively leveraging both localized and contextual information. The ViT/VideoMAE design lacks this hybridization, limiting its capacity to capture spatial hierarchies.

4.8.3 Correlation with MELD Dataset

Single-Frame Facial Processing: MELD frequently offers static or intermittently sampled facial data. ConvNeXtEnhancedFRAM effectively handles this via single-frame processing. ViT/VideoMAE, on the other hand, depends on temporal sequences and sometimes uses zero-padding for missing frames, which introduces noise and degrades performance.

Resilience to Data Inconsistencies: ConvNeXtEnhancedFRAM is more adept at handling variations in illumination, occlusion, and resolution due to its convolutional design and focus on salient facial regions. The global attention mechanism of ViT/VideoMAE is more sensitive to noise and less effective in the presence of inconsistent data.

Addressing Class Imbalance: ConvNeXtEnhancedFRAM employs weighted sampling to address the underrepresentation of certain emotion classes (e.g., fear, disgust), thereby improving classification performance across all labels. The ViT/VideoMAE architecture lacks this capability, potentially skewing predictions toward majority classes.

4.8.4 Advantages of Training and Optimization

Gradient Accumulation: ConvNeXtEnhancedFRAM uses gradient accumulation to emulate larger batch sizes and stabilize training within GPU constraints. This results in smoother convergence. ViT/VideoMAE does not employ this technique, which may affect training stability.

Simplified Loss Function: ConvNeXtEnhancedFRAM focuses solely on emotion classification, avoiding the complexities of auxiliary tasks. In contrast, ViT/VideoMAE

includes Action Unit (AU) loss, which can introduce noise from missing AU labels in the MELD dataset.

Selective Fine-Tuning: `ConvNeXtEnhancedFRAM` strategically unfreezes certain layers for fine-tuning, maintaining a balance between generalization and task-specific adaptation. `ViT/VideoMAE` fine-tunes only a few layers and lacks convolutional adaptability.

4.8.5 Empirical Efficacy

Experimental results show that `ConvNeXtEnhancedFRAM` achieves superior fusion accuracy compared to `ViT/VideoMAE` architectures. The use of region-based attention, convolutional hierarchies, and optimized training strategies directly contributes to its improved performance. In contrast, the `ViT/VideoMAE` approach is more sensitive to the MELD dataset's inconsistencies and lacks architectural enhancements that focus on emotion-relevant visual areas.

4.8.6 Conclusion

The `ConvNeXtEnhancedFRAM` model outperforms the `ViT/VideoMAE` architecture in several aspects, including spatial modeling, attention targeting, training stability, and dataset alignment. Its convolutional backbone, region-focused facial attention, and data-informed design make it well-suited for emotion recognition tasks on MELD. Empirical results reinforce these strengths, confirming `ConvNeXtEnhancedFRAM` as a superior architecture for multimodal emotion analysis.

4.9 Comparison with State of the Art Model

Model Configuration	Modalities Used	Accuracy (%)	Remarks
State of the Art [30]	Text + Face + Video + Voice	66.36	Reference model for comparison
ViT/VideoMAE + Action Units (Ours)	Text + Face + Video + Voice	39.83	Incorporates AU loss; less robust to missing or inconsistent data
ConvNeXtEnhanced (Ours)	Text + Face + Video + Voice	61.12	Employs convolutional hierarchies, region-based attention, and training enhancements

Table 4.8: Performance Comparison Between Models Using Text, Face, Video, and Voice Modalities

Chapter 5

Discussion

5.1 Discussion

The development of a robust multimodal emotion detection system has yielded significant insights—both technical and intuitive—regarding the contribution of various data sources to the comprehension of human emotions.

A key result of this research was the comparative assessment between the baseline and our suggested models. The baseline model, integrating text, facial recognition, video, and audio modalities, attained the best accuracy of **66.36%**. This result underscores the significance of textual data in emotion identification. Language carries profound conceptual and emotional nuances that are sometimes difficult to convey solely through nonverbal cues. The lack of textual input in the proposed models resulted in a significant decline in performance throughout our testing. The *ConvNeXtEnhanced* model, the most effective in our pipeline, attained an accuracy of **61.12%**, which, while encouraging, did not surpass the baseline. The performance disparity primarily reflects the emotional subtleties that are lost when textual data is absent.

A significant discovery that arose during model assessment was the inconsistency of the facial modality. Notwithstanding the use of a specialized Action Unit (AU) classifier and established frame-level models for facial analysis, the contributions of the facial data were inconsistent. In several real-world instances, factors such as inadequate illumination, occluded faces, extreme head orientations, or low-resolution images hindered the model's

ability to extract meaningful features. In these instances, the model resorted to padding the input, which introduced noise rather than useful signals. Consequently, the fusion process often reduced the influence of facial information, placing more emphasis on audio or video data.

It is important to emphasize the significance of our fusion technique in addressing these constraints. The cross-attention fusion mechanism demonstrated efficacy in acquiring context-aware weighting of modalities. When a certain stream, such as facial data, was compromised, the model instinctively relied more on dependable streams like audio or video. This behavior highlights a core feature of our architecture: its capacity to prioritize stronger signals over weaker or noisy ones. This adaptability helps mitigate some issues associated with fluctuating data quality, although it also reveals the model's susceptibility to input imbalance.

The considerable disparity between the two proposed models emphasizes the influence of architectural decisions. The *ViT/VideoMAE + AU* model, although using comparable modalities, exhibited markedly lower performance (**39.83%** accuracy) relative to the *ConvNeXtEnhanced* variant. This shortfall can be attributed to the absence of hierarchical spatial processing and region-based attention mechanisms in the former. In contrast, the *ConvNeXtEnhanced* model incorporated structural enhancements that improved comprehension of localized features in visual inputs and facilitated a more stable training process.

This research demonstrates that effective emotion identification requires more than merely aggregating multiple modalities. It demands careful architectural design, prudent handling of variable input data, and a deep understanding of the diverse ways in which humans express emotion across different contexts and media. The models we developed—and their resultant outcomes—offer valuable insight into the intricate relationship between human behavior and machine interpretation. Despite several challenges, the findings lay a foundation for the development of more advanced and context-aware emotion recognition systems.

Chapter 6

Conclusion And Future Work

6.1 Conclusion

This research established a multimodal emotion recognition system that employs audio, video, and facial data to categorize emotions such as neutral, anger, contempt, fear, joy, sorrow, and surprise. The model used advanced architectures, including Wav2Vec2 for audio processing, Vision Transformer (ViT) for video, and VideoMAE for facial sequence analysis, interconnected via a cross-attention fusion method. The model included ConvNeXt with an Enhanced Feature Region Attention Module (FRAM) in both the VideoEncoder and FaceEncoder, improving feature extraction by concentrating on geographically and temporally significant areas in video and facial frames. The ConvNeXt integration enhanced the model's capacity to discern intricate visual details, including subtle facial expressions and contextual scene components, hence augmenting its efficacy in emotion recognition. The system exhibited adequate performance; nevertheless, a significant restriction was the lack of text data, which negatively affected accuracy. Literature and our tests affirm that language cues provide substantial emotional and semantic context, facilitating more accurate emotion comprehension. The absence of text input hindered the model's ability to attain similar performance, highlighting the significance of multimodal integration.

A significant discovery was the variable impact of the facial modality. Notwithstanding the use of a specialized Action Unit (AU) classifier and VideoMAE with ConvNeXt

for facial frame analysis, the quality of facial data exhibited considerable variability across samples. Poor lighting, occlusion, and improper camera angles often hindered the extraction of significant face characteristics. In these instances, the system used zero tensor padding, which introduced noise that decreased the effectiveness of the fusion process. As a result, the fusion module often diminished the face modality's significance, placing more emphasis on audio and video streams, where ConvNeXt's spatial attention improved resilience. These discoveries underscore the need for enhanced facial data pretreatment and effective management of changing input quality to optimize the capabilities of ConvNeXt and other components, facilitating future advancements in multimodal emotion identification systems.

6.2 Future Work

There are numerous potential avenues for expansion in the future of this study. The incorporation of temperature as an additional modality is a promising direction. According to research, ambient temperature can either influence or indicate emotional states. For example, cooler environments (0–10°C) are frequently associated with feelings of melancholy or lethargy, while higher temps (30–40°C) are linked to exhilaration, tension, or increased arousal. Francisco Barbosa Escobar and Qian Janice Wang's "The Temperature of Emotions" emphasizes the correlation between emotional experience and physical temperature. Particularly in outdoor or real-world scenarios where internal sentiments intersect with external conditions, the integration of such environmental data into emotion recognition systems may provide valuable contextual indicators.

Another substantial improvement would entail enhancing the model's resilience in the presence of facial data that is either absent or of low quality. In order to reduce the influence of incomplete inputs, future models may implement methodologies such as confidence-weighted fusion, attention-based masking, or feature imputation.

Additionally, the system's expansion to accommodate multiparty interactions, in which conversations involve multiple participants and affective expressions fluctuate in accordance with intricate social dynamics, presents an additional opportunity. Such functionality could enhance the model's suitability for group contexts, including therapy ses-

sions, meetings, and collaborative virtual environments.

Furthermore, haptic feedback is a promising future integration. Haptic modalities provide a tactile medium for the transmission or perception of emotional information. For instance, emotional states may be reflected or influenced through the use of mild vibrations, temperature changes, or pressure-based feedback. The interactive experience may be enhanced by integrating haptic signals into emotion recognition or response systems, particularly in immersive contexts like virtual or augmented reality.

Subsequently, the concept has substantial potential for integration into virtual reality (VR) environments. A rich, immersive environment is provided by VR, which enables the dynamic analysis and response to emotional expressions. By incorporating emotion recognition into VR systems, we could improve the functionality of applications such as adaptive narration, therapeutic simulations, and emotionally responsive virtual agents. This would contribute to the advancement of our comprehension and modeling of human affect in a variety of interactive contexts.

Bibliography

[1] Ekman, P. (1992). An Argument for Basic Emotions. Cognition Emotion, 6(3-4), 169–200.

[2] Plutchik, R. (1980). A General Psychoevolutionary Theory of Emotion. In R. Plutchik H. Kellerman (Eds.), Emotion: Theory, Research, and Experience (Vol. 1, pp. 3–33). Academic Press.

[3] Mehrabian, A. (1971). Silent Messages: Implicit Communication of Emotions and Attitudes. Belmont, CA: Wadsworth.

[4] Damasio, A. R. (1994). Descartes' Error: Emotion, Reason, and the Human Brain. New York: G.P. Putnam's Sons.

[5] Ekman, P., Friesen, W. V. (1978). Facial Action Coding System (FACS). Consulting Psychologists Press..

[6] Martinez, B., Valstar, M. F. (2016). Advances in automatic facial expression recognition. Springer, pp. 63–100.

[7] Schuller, B., Steidl, S., Batliner, A. (2009). The INTERSPEECH 2009 Emotion Challenge. INTERSPEECH.

[8] Eyben, F., Wöllmer, M., Schuller, B. (2010). openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor. ACM Multimedia.

[9] Mohammad, S. M., Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. Computational Intelligence, 29(3), 436–465.

[10] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.

[11] Zeng, Z., Pantic, M., Roisman, G. I., Huang, T. S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(1), 39–58.

[12] Schuller, B., Batliner, A., Steidl, S., Seppi, D. (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Communication, 53(9–10), 1062–1087.

[13] Wieser, M. J., Brosch, T. (2012). Faces in context: A review and systematization of contextual influences on affective face processing. Frontiers in Psychology, 3, 471.

[14] Lian, H., Lu, C., Li, S., Zhao, Y., Tang, C., Zong, Y. (2023). A Survey of Deep Learning-Based Multimodal Emotion Recognition: Speech, Text, and Face. Entropy, 25(10), 1440.

[15] Khalane, M., et al. (2025). Evaluating significant features in context-aware multimodal emotion recognition with XAI methods. Expert Systems, Wiley Online Library.

[16] Palash, M., Bhargava, B. (2023). EMERSK – Explainable Multimodal Emotion Recognition with Situational Knowledge. arXiv preprint arXiv:2306.08657.

[17] Kahou, S. E., et al. (2015). EmoNets: Multimodal deep learning approaches for emotion recognition in video. arXiv preprint arXiv:1503.01800.

[18] Abbaschian, S., Sierra-Sosa, D., Elmaghraby, A. (2021). Deep learning techniques for speech emotion recognition, from databases to models. Sensors, 21(4), 1249.

[19] Bharti, S. K., Chaudhary, V., Ghosh, S., Bhattacharyya, D. (2022). Hybrid deep learning approach for emotion classification from text. Computational Intelligence and Neuroscience, 2022, 1–12.

[20] Latif, S., Rana, R., Khalifa, S., Jurdak, R., Schuller, B. (2018). Transfer learning for improving speech emotion classification accuracy. arXiv preprint arXiv:1801.06353.

[21] Peng, J., Wang, Z., Lu, J. (2021). A multi-scale CNN model with attention for speech emotion recognition. arXiv preprint arXiv:2106.04133.

[22] Saeed, S., Aslam, M. W., Javaid, N. (2024). A comprehensive review on deep learning-based facial emotion recognition systems. Sensors, 24(11), 3484.

[23] Zhang, Z., Fort, J. M., Giménez Mateu, L. (2023). Frontiers in Psychology, 14, 1280136.

[24] Indrasiri, P. L., et al. (2024). arXiv preprint arXiv:2412.02283.

[25] Yu, Y., et al. (2024). Biomedical Signal Processing and Control, 86, 104398.

[26] Agarwal, S., Mukesh, D., Gupta, K. EMOTION DETECTION USING CONTEXT BASED FEATURES USING DEEP LEARNING TECHNIQUE.

[27] Hertenstein MJ, Keltner D, App B, Bulleit BA, Jaskolka AR. Touch communicates distinct emotions. Emotion. 2006 Aug;6(3):528-33. doi: 10.1037/1528-3542.6.3.528. PMID: 16938094.

[28] Wang, Rongrong Quek, Francis. (2010). Touch talk: Contextualizing remote touch for affective interaction. 13-20. 10.1145/1709886.1709891.

[29] Barbosa Escobar, Francisco Velasco, Carlos Motoki, Kosuke Byrne, Derek Wang, Qian. (2021). The temperature of emotions. PLOS ONE. 16. e0252408. 10.1371/journal.pone.0252408.

[30] Farhadipour, Aref Ranjbar, Hossein Chapariniya, Masoumeh Vukovic, Teodora Ebling, Sarah Dellwo, Volker. (2025). Multimodal Emotion Recognition and Sentiment Analysis in Multi-Party Conversation Contexts. 10.48550/arXiv.2503.06805.

[31] Sandratra Rasendrasoa. Multimodal and multi-party interactions in mixed communities. Artificial Intelligence [cs.AI]. Normandie Université, 2023. English. NNT: 2023NORMR086. tel-04705429.

[32] Yu, Di Fu, Haiyue Song, Yanchen Xie, Wenjian Zhijie, Xie. (2023). Deep transfer learning rolling bearing fault diagnosis method based on convolutional neural network feature fusion. Measurement Science and Technology. 35. 10.1088/1361-6501/acfe31.

[33] Fabien Boucaud, Catherine Pelachaud, and Indira Thouvenin. 2021. Decision Model for a Virtual Agent that can Touch and be Touched. In Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '21). International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 232–241.

[34] Boucaud, Fabien Tafiani, Quentin Pelachaud, Catherine Thouvenin, Indira. (2019). Social Touch in Human-agent Interactions in an Immersive Virtual Environment. 129-136. 10.5220/0007397001290136.

[35] Hauser, Steven McIntyre, Sarah Israr, Ali Olausson, Håkan Gerling, Gregory. (2019). Uncovering Human-to-Human Physical Interactions that Underlie Emotional and Affective Touch Communication. 2019 IEEE World Haptics Conference (WHC). World Haptics Conference (2019 : Tokyo, Japan). 2019. 10.1109/WHC.2019.8816169.